2020

Essays in program evaluation

Niklaus Carlton Julius
Iowa State University

Essays in program evaluation

by

Niklaus Carlton Julius

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:
Helle Bunzel, Co-major Professor
Otávio Bartalotti, Co-major Professor
Brent Kreider
Ulrike Genschel
Désiré Kédagni

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank everyone who supported me in the course of completing this work. First and foremost, my co-major professors, Dr. Helle Bunzel and Dr. Otávio Bartalotti, for their guidance, patience, and unwavering support throughout my time as a graduate student. I also thank Dr. Gray Calhoun and Dr. Brent Kreider for their invaluable advice and support in the initial stages of my graduate career. I thank my committee members, Dr. Ulrike Genschel and Dr. Désiré Kédagni, for their efforts and flexibility. Finally, I would like to thank the members of the Econometrics Reading Group for their insightful input on parts of this work.

ABSTRACT

This dissertation consists of three chapters on program evaluation, or the estimation of treatment effects.

The first chapter discusses bootstrap methods for inference on matching estimators, a popular approach to program evaluation. Abadie and Imbens (2008) showed that the standard non-parametric bootstrap fails to provide valid inference with matching estimators, and conjectured that a wild bootstrap could solve the problem. Otsu and Rai (2017) confirmed this conjecture, providing a wild bootstrap procedure that is valid in general. Their bootstrap builds in a bias correction procedure that requires estimation of conditional mean functions, a procedure that is generally necessary for consistent matching estimation. However, this step also introduces a new source of estimation error, lessening the efficiency of the bootstrap. I show that even in a special case, when bias correction in the estimator is unnecessary, the conditional mean function estimation is a required element of any wild bootstrap for the matching estimator. This shows that the Otsu and Rai bootstrap cannot be modified to be more efficient even by leveraging much stronger assumptions. Simulations provide additional support for this conclusion.

The second chapter also deals with matching estimators. I consider the problem faced by a practitioner who wishes to use matching estimation to estimate a treatment effect - in particular, choosing from a large set of available matching procedures. I cast matching estimators as two-step procedures - a weight-generation step followed by a weighted difference in means - and derive weights that minimize mean-squared error (MSE) under certain conditions. Understanding why the optimal weights behave the way they do generates insights about which matching procedures are likely to minimize MSE, enabling practitioners to use their economic intuition, knowledge of the empirical context, and knowledge of the sam-

pling process to choose an appropriate matching procedure. I develop a simple 'augmented' matching procedure to illustrate, and through simulation confirm that the guidance I offer is correct.

In the final chapter, I apply my program evaluation expertise to a question in the economics of education - specifically, the effect of teacher gender on student test scores. Previous literature in this vein has focused on the estimation of average effects. By exploiting random assignment of students to teachers in a field experiment, I study heterogeneity in the impact of teacher gender on math and reading test outcomes for primary school students of differing ability. I find that assignment to a female teacher is generally positive for male students, while it has no significant effect for female students. In addition, I find very little heterogeneity in the effect of teacher gender along the ability axis, suggesting that average effect estimates from previous investigations do not mask significant heterogeneity. My results are consistent with differential teacher behavior based on gender stereotypes, and somewhat inconsistent with differential student behavior based on gender stereotypes.

CHAPTER 1. A NOTE ON BOOTSTRAPS FOR MATCHING ESTIMATION

Matching estimators are a popular approach to program evaluation. Abadie and Imbens (2008) showed that the naive bootstrap fails to provide valid inference with matching estimators, and conjectured that a wild bootstrap could solve the problem. Otsu and Rai (2017) confirm this conjecture. I show that even with much stronger assumptions, the Otsu and Rai (2017) bootstrap cannot be modified to be more efficient.

1.1 Introduction

Evaluating the efficacy of programs or treatments requires the estimation of treatment effects. A popular nonparametric method for estimating average treatment effects is the method of matching, which has significant intuitive appeal. These methods match treated units to control units that are 'close' as measured by a chosen metric. The estimated average treatment effect is then constructed by averaging the differences between matched units.

Matching can be done with or without replacement, but the latter is more common. Abadie and Imbens (2006) began a comprehensive study of matching estimators, continued in a series of papers (Abadie and Imbens, 2008, 2011, 2009, 2016). Abadie and Imbens (2006) found that matching on covariates is not always $\sqrt{N}$-consistent, generally requiring a bias correction. Abadie and Imbens (2008) showed that even when the matching estimator is $\sqrt{N}$-consistent, the 'naive' bootstrap[1] fails to correctly estimate the distribution of the matching estimator.

[1] Resampling observations with replacement to create a bootstrapped sample.

Abadie and Imbens (2008) traced the failure of the naive bootstrap to a failure to capture the behavior of the matching process that underlies the estimator. Specifically, a given control observation will tend not to be matched to the same treated units, or even the same number of treated units, across the true and bootstrapped samples. Abadie and Imbens (2008) noted that this reasoning clearly suggests a wild bootstrap could avoid the problem, by conditioning the bootstrap on realized matches in the true sample. Otsu and Rai (2017) confirm this conjecture, providing a consistent bootstrap procedure for matching estimators that match on covariates.

Because Otsu and Rai (2017) developed a bootstrap that is valid in general, it naturally incorporates the bias correction that is sometimes required for consistency of the matching estimator. This entails the estimation of conditional mean functions for units in both treatment arms, which introduces an additional source of estimation error. It is reasonable to suspect that eliminating this estimation error might generate efficiency gains in the bootstrap, at the cost of generality.

In this chapter, I show that conditional mean estimation is necessary for a valid wild bootstrap even without bias correction. Considering a special case where matching estimation is consistent without bias correction, I develop a natural wild bootstrap and show that it fails to consistently estimate the variance of the matching estimator. Potential solutions to the problem fall into two categories: those that essentially reproduce the bootstrap from Otsu and Rai (2017), and those that abandon the wild bootstrap entirely.

The remainder of the chapter is organized as follows. In Section 1.2, I introduce notation and give a formal explanation of matching estimators. In Section 1.3, I propose a wild bootstrap without bias correction, and show theoretically that it fails in general. In Section 1.4, I provide simulation evidence of the failure. Finally, in Section 1.5 I conclude by providing an intuition for the failure and some potential avenues for future research.

1.2 Setup and Notation

Suppose we observe a random sample of size $N = N_1 + N_0$, which consists of N_1 units that received treatment and N_0 units that did not receive treatment. For each unit $i = 1, ..., N$, we observe a triplet consisting of a treatment indicator $D_i \in \{0, 1\}$, a covariate X_i, and the outcome variable $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for unit i when $D_i = 1$ and $D_i = 0$, respectively.

In general, X_i can be a vector of multiple covariates. In this chapter, I restrict my attention to the case where X_i is scalar, as this is a simple way to eliminate the need for bias correction. While it is possible for X_i to be a vector of multiple covariates *and* for bias correction to be unnecessary, it has no impact on my conclusions whether X_i is a scalar or vector.

Given this sample, we seek to conduct inference on the average treatment effect for the treated population[2] (the ATT),

$$\tau^t = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right] \tag{1.1}$$

To estimate τ^t, we use an M nearest-neighbor matching estimator of the type studied in Abadie and Imbens (2006). Matching is based on covariate distance. Formally, the estimator is described as follows:

$$\hat{\tau}^t = \frac{1}{N_1} \sum_{i:D_i=1} \left(Y_i(1) - \widehat{Y_i(0)} \right) \tag{1.2}$$

where $\widehat{Y_i(0)}$ is an estimate of the unobserved potential outcome, defined as

$$\widehat{Y_i(0)} = \begin{cases} Y_i, & \text{if } D_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j, & \text{if } D_i = 1 \end{cases} \tag{1.3}$$

[2]The restriction to the ATT is without loss of generality. The extension to the case of the average treatment effect for the untreated population (the ATC) is straightforward, and extending the result to the case of the average treatment effect for the whole population (the ATE) follows from the representation of the ATE as a weighted average of the ATT and the ATC.

$\mathcal{J}_M(i)$ is the set of indices describing the M closest matches to unit i. Formally, it is defined as

$$\mathcal{J}_M(i) = \left\{ j \in \{1, ..., N\} : D_j = 0, \sum_{l:D_l=0} \mathbb{I}\{ \mid X_l - X_i \mid \leq \mid X_j - X_i \mid \} \leq M \right\} \tag{1.4}$$

As an example, suppose that for some treated unit i, $X_i = 4$. Suppose there are three control units j, k, l, with $X_j = 3.5$, $X_k = 4.6$, $X_l = 5$. $\mathcal{J}_1(i)$ is the closest match, and would thus be the singleton set $\{j\}$. $\mathcal{J}_2(i)$ would be $\{j, k\}$, and $\mathcal{J}_3(i)$ would be $\{j, k, l\}$. For the remainder of the chapter, I generally restrict attention to the case where $M = 1$ - a common choice in practice, and one which eases exposition considerably.

It is useful to define K_i as the number of times unit i is used as a match

$$K_i = \begin{cases} 0, & \text{if } D_i = 1, \\ \sum_{j:D_j=1} \mathbb{I}\{i \in \mathcal{J}_M(i)\}, & \text{if } D_i = 0 \end{cases} \tag{1.5}$$

Let $m(i)$ be a function that returns the single value in $\mathcal{J}_M(i)$ when $M = 1$. Finally, let $\mu(d, x) = \mathbb{E}[Y \mid D = d, X = x]$ and $\sigma^2(d, x) = \text{Var}(Y \mid D = d, X = x)$.

Abadie and Imbens (2006), in their study of matching estimators of this kind, considered the case where N grows while M remains constant. Otsu and Rai (2017) refer to this as 'fixed-M asymptotics'. Under the following assumptions, Abadie and Imbens were able to characterize the asymptotic behavior of $\hat{\tau}^t$.

AI.1 *Conditional on $D_i = d$, the sample consists of independent draws from $Y, X \mid D = d$ for $d \in \{0, 1\}$. For some $r \leq 1$, $N_1^r / N_0 \to \theta \in (0, \infty)$.*

AI.2 *X is continuously distributed on compact and convex support $\mathbb{X} \subset \mathbb{R}$. The density of X is bounded and bounded away from zero on $\mathbb{X}$.*

AI.3 *D is independent of $Y(0)$ conditional on $X = x$ for almost every x. There exists a positive constant c such that $Pr[D = 1 \mid X = x] \leq 1 - c$ for almost every x.*

AI.4 *For $d \in \{0, 1\}$, $\mu(d, x)$ and $\sigma^2(d, x)$ are Lipschitz in $\mathbb{X}$, and $\sigma^2(d, x)$ is bounded away from zero on $\mathbb{X}$.*

Assumptions AI.1 through AI.3 are relatively standard. They provide useful conditions on the sampling process and the distribution of X, along with the standard unconfoundedness and overlap assumptions necessary to identify the ATT. AI.4 provides smoothness and bounds that are necessary for the characterization of the bias term and its asymptotic behavior.

Under these assumptions, Abadie and Imbens (2006) showed that $\hat{\tau}^t \to^p \tau^t$, and

$$\frac{\sqrt{N_1}\left(\hat{\tau}^t - B_N^t - \tau^t\right)}{\sigma_N^t} \to^d \mathcal{N}(0,1) \tag{1.6}$$

where

$$\begin{aligned}
B_N^t &= \sum_{i=1}^{N} D_i \left(\frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} [\mu(0, X_i) - \mu(0, X_j)] \right) \\
(\sigma_N^t)^2 &= (\sigma_{1N}^t)^2 + (\sigma_2^t)^2 \\
\sigma_{1N}^t &= \frac{1}{N_1} \sum_{i=1}^{N} \left(D_i + (1 - D_i) \frac{1}{M} K_i \right)^2 \sigma^2(D_i, X_i) \\
(\sigma_2^2)^2 &= \mathbb{E}\left[(\mu(1, X_i) - \mu(0, X_i) - \tau^t)^2 \mid D_i = 1 \right]
\end{aligned} \tag{1.7}$$

Inuitively, B_N^t captures the bias produced by matches that are less than perfect. Since selection on observables is contained in assumptions AI.1 through AI.4, if $X_j = X_i$, $\mu(0, X_i) - \mu(0, X_j) = 0$ follows trivially. σ_{1N}^t captures the variance effect of units being matched potentially multiple times[3]. Finally, $(\sigma_2^t)^2$ captures the variance produced by innate differences in the treatment effect conditional on X. If τ^t did not depend on X, for instance, $(\sigma_2^t)^2$ would trivially be zero.

When X_i is scalar, B_N^t is $o_p(N_1^{-1/2})$ and thus asymptotically ignorable. Alternatively, I could directly restrict attention to cases where B_N^t is $o_p(N_1^{-1/2})$, which occurs when $r > k/2$, effectively placing a strong restriction on the growth rate of the different treatment arms in the sampling process. Intuitively, B_N^t is ignorable when N_0 grows at least as fast as $\sqrt{N_1}$. It is more common for control groups to be at least as large as treatment groups, so for

[3]Note that if $K_i = 1$ for all i, σ_{1N}^t devoles into the average of $\sigma^2(D_i, X_i)$ over the sample.

the purposes of estimating the ATT this requirement is often satisfied, further supporting an exploration for a special case of the bootstrap without bias correction. When B_N^t is not ignorable, the bias term grows because the number of matches required grows faster than the 'quality'[4] of matches shrinks, resulting in the average 'quality' of matches decreasing as N increases.

1.3 Proposed Bootstrap

Otsu and Rai (2017) provide a valid wild bootstrap for the general case of the M-nearest neighbor matching estimator described above, for any number of continuous covariates. In their setting, B_N^t is not guaranteed to be $o_p(N_1^{-1/2})$. Thus, as a bias correction is needed in the estimation procedure, part of the bootstrap must also account for the variance generated by the bias correction. Otsu and Rai solve this problem by simply bootstrapping the bias correction itself, which requires estimating of the conditional mean functions $\mu(d, x)$ for $d \in \{0, 1\}$. It is reasonable to suspect that estimation errors in this step may increase the estimated variance of $\hat{\tau}^t$, and thus that a bootstrap without the bias correction might be more efficient, at the cost of being invalid when a bias correction is needed.

The reason a wild bootstrap was conjectured by Abadie and Imbens (2008) is that by definition, a wild bootstrap will not change the matches and thus will not need to estimate the distribution of K_i. I consider the following procedure:

1. Estimate $\hat{\tau}^t$ using nearest-neighbor matching.
2. Using $\hat{\tau}^t$ from step 1, generate residuals $\hat{\xi}_i = \left(Y_i(1) - \hat{\tau}^t\right) - \widehat{Y_i(0)}$.
3. Draw a bootstrap auxiliary variable ϵ_i^* from the Rademacher distribution.
4. Create bootstrapped treated outcomes $Y_i(1)^* = \widehat{Y_i(0)} + \hat{\tau}^t + \epsilon_i^* \hat{\xi}_i$.
5. Estimate $\hat{\tau}^{t*}$ using nearest-neighbor matching on the bootstrapped sample.

[4] The 'quality' of a match can be thought of as the difference between $\mu(X_i, 0)$ and $\mu(X_j, 0)$ for i and j being matched together.

6. Repeat steps 1-5 B times, and use the sample variance of $\hat{\tau}^{t*}$ to estimate the variance of $\hat{\tau}^t$.

Davidson et al. (2007) provides strong evidence that the Rademacher distribution is superior for wild bootstrap performance. Indeed, their results suggest that the Rademacher distribution is one of the best distributions possible. The Rademacher distribution is very simple, with ϵ_i^* taking the values 1 and -1 with equal probability.

This is a *prima facie* reasonable procedure. While it may appear odd at first for the bootstrapped dataset to consist of $(Y(1)^*, Y(0))$, this is a result of estimating the ATT. Bootstrapping $Y(0)$ would require defining and constructing estimates of $\widehat{Y_i(1)}$, which cannot be done without estimating the average treatment effect for the whole population, or estimating conditional mean functions as in Otsu and Rai (2017). As τ^t is often an object of interest in itself, a wild bootstrap for this case is worth having.

Unfortunately, the proposed bootstrap is not valid in general. Furthermore, the failure indicates that any wild bootstrap that does not construct $\widehat{Y_i(1)}$ will not be valid in general. The failure is not total - in certain special cases[5], the procedure works correctly. The special cases offer an intuition for why the procedure fails in general. Solving this problem essentially requires replicating the Otsu and Rai (2017) bootstrap procedure by estimating conditional mean functions to construct $\widehat{Y_i(1)}$, or abandoning the wild bootstrap altogether.

The simplest option in the latter category is to bootstrap the treatment indicator D_i rather than Y_i. As this approach relies on estimating propensity scores, which is done for all observations even when estimating the ATT, it avoids the incomplete bootstrapping issue. This approach is illustrated in Huber et al. (2016) and Adusumilli (2017).

For the proposed bootstrap to work, it would suffice for the following to be true,

$$\sup_q \left| \Pr\left\{ \sqrt{N_1}\left(\hat{\tau}^{t*} - \hat{\tau}^t\right) \leq q \mid \mathbf{Z} \right\} - \Pr\left\{ \sqrt{N_1}\left(\hat{\tau}^t - \tau^t\right) \leq q \right\} \right| \to^p 0 \tag{1.8}$$

[5]Most notably, in the case where each treated unit has a *unique* closest match in the control group, and also in the case where treated units have zero idiosyncratic errors.

where $\mathbf{Z} = \{Y, D, X\}$ is the entire sample. Abadie and Imbens (2006) show (in Corollary 1) that $\sqrt{N_1}(\hat{\tau}^t - \tau^t)/\sigma_N^t$ is asymptotically normal, so (1.8) implies the following:

$$\text{Var}\left(\sqrt{N_1}\left(\hat{\tau}^{t*} - \hat{\tau}^t\right) \mid \mathbf{Z}\right) - (\sigma_N^t)^2 \to^p 0 \tag{1.9}$$

$$\left|\Pr\left\{\sqrt{N_1}\left(\hat{\tau}^{t*} - \hat{\tau}^t\right)/\sigma_N^t \leq t \mid \mathbf{Z}\right\} - \Phi(t)\right| \to^p 0 \quad \forall\, t \in \mathbb{R} \tag{1.10}$$

Intuitively, (1.9) requires that the bootstrapped estimates $\hat{\tau}^{t*}$ have the correct variance, and (1.10) requires that the bootstrapped estimates are asymptotically normally distributed. Note that *either* condition failing to hold is sufficient to prove that the bootstrap does not work. I will give an abbreviated proof here, as the failure is interesting.

It is possible to recover a representation for $\hat{\tau}^{t*}$ from the proposed bootstrap procedure,

$$\hat{\tau}^{t*} = \hat{\tau}^t + \frac{1}{N_1}\sum_{i=1}^{N} D_i \hat{\xi}_i \epsilon_i^* \tag{1.11}$$

This representation can be decomposed into a form involving estimated population parameters and the parameters themselves:

$$\begin{aligned}\hat{\tau}^{t*} &= \hat{\tau}^t + \frac{1}{N_1}\sum_{i=1}^{N} D_i \xi_i \epsilon_i^* + \frac{1}{N_1}\sum_{i=1}^{N} D_i\left(\hat{\xi}_i - \xi_i\right)\epsilon_i^* \\ &= \hat{\tau}^t + T_N^{t*} + Q_N^{t*} + R_N^{t*}\end{aligned} \tag{1.12}$$

where

$$\begin{aligned}T_N^{t*} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i\left(\mu(1, X_i) - \mu(0, X_i) - \tau^t\right)\epsilon_i^* \\ Q_N^{t*} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i\left(Y_i(1) - \widehat{Y_i(0)} - \mu(1, X_i) + \mu(0, X_i)\right)\epsilon_i^* \\ R_N^{t*} &= \frac{1}{N_1}\sum_{i=1}^{N} D_i\left(\tau^t - \hat{\tau}^t\right)\epsilon_i^*\end{aligned} \tag{1.13}$$

In simple terms, T_N^{t*} is a term capturing the differences between the true treatment effect at some value of X_i and the ATT. Q_N^{t*} captures the variance contributed by the 'quality' of the matches as well as some remainder terms, and R_N^{t*} is a pure remainder term.

Noting that $\sqrt{N_1}\left(\hat{\tau}^{t*} - \hat{\tau}^t\right) = \sqrt{N_1}\left(T_N^{t*} + Q_N^{t*} + R_N^{t*}\right)$, it follows that

$$\begin{aligned}\text{Var}\left(\sqrt{N_1}\left(\hat{\tau}^{t*} - \hat{\tau}^t\right) \mid \mathbf{Z}\right) &= N_1\mathbb{E}\left[(T_N^{t*})^2 + (Q_N^{t*})^2 + (R_N^{t*})^2 \mid \mathbf{Z}\right] \\ &\quad + N_1\mathbb{E}\left[2\left(T_N^{t*}Q_N^{t*} + T_N^{t*}R_N^{t*} + Q_N^{t*}R_N^{t*}\right) \mid \mathbf{Z}\right] \qquad (1.14)\end{aligned}$$

Under assumptions A1-A4, the following results hold:

$$\begin{aligned}\mathbb{E}\left[N_1(T_N^{t*})^2 \mid \mathbf{Z}\right] &\to^p (\sigma_2^t)^2 \\ \mathbb{E}\left[N_1(Q_N^{t*})^2 \mid \mathbf{Z}\right] &\to^p (\sigma_{1N}^t)' \\ \mathbb{E}\left[N_1(R_N^{t*})^2 \mid \mathbf{Z}\right] &\text{ is } O_p(N_1^{-1/2}) \\ \mathbb{E}\left[2N_1\left(T_N^{t*}Q_N^{t*} + T_N^{t*}R_N^{t*} + Q_N^{t*}R_N^{t*}\right) \mid \mathbf{Z}\right] &= 0 \qquad (1.15)\end{aligned}$$

The full proof is relegated to Appendix A. Note that $\mathbb{E}\left[N_1(Q_N^{t*})^2 \mid \mathbf{Z}\right]$ does *not* converge to $(\sigma_{1N}^t)^2$. This is the failure of the bootstrap procedure. Instead, $\mathbb{E}\left[N_1(Q_N^{t*})^2 \mid \mathbf{Z}\right]$ converges to $(\sigma_{1N}^t)'$,

$$(\sigma_{1N}^t)' = \frac{1}{N_1}\sum_{i=1}^{N}\left(D_i + (1 - D_i)\frac{1}{M}K_i\right)\sigma^2(D_i, X_i) \qquad (1.16)$$

When contrasted with $(\sigma_{1N}^t)^2$, the term involving K_i lacks a power of two. It is easy to see from this why the bootstrap works when each treated unit has a unique matching control unit - in that case, $K_i = 1$ for all i, so the missing power of two has no effect and $(\sigma_{1N}^t)' = (\sigma_{1N}^t)^2$. Due to this failure, the variance of the bootstrapped $\hat{\tau}^{t*}$'s is incorrect, and thus the bootstrap consistency condition does not hold.

1.4 Simulations

In this section, I use the data generating process from Abadie and Imbens (2008), which is described as follows:

1. The marginal distribution of X is uniform on the interval $[0, 1]$.

2. The ratio of treated units to control units is $N_1/N_0 = \alpha$ for some positive α.

3. The propensity score $e(X) = \Pr[D = 1 \mid X = x]$ is a constant function of α.

4. The distribution of $Y(1)$ is degenerate, with $\Pr[Y_i(1) = \tau] = 1$.

5. The conditional distribution of $Y(0) \mid X = x$ is standard normal.

This DGP enabled Abadie and Imbens (2008) to find an analytic representation for both the conditional and unconditional variance of $\hat{\tau}^t$,

$$\begin{aligned} \text{Var}\left(\hat{\tau}^t\right) &= \frac{1}{N_1} + \frac{3}{2}\frac{(N_1 - 1)(N_0 + 8/3)}{N_1(N_0 + 1)(N_0 + 2)} \\ \text{Var}\left(\hat{\tau}^t \mid \mathbf{Z}\right) &= \frac{1}{N_1^2}\sum_{i=1}^{N} K_i^2 \end{aligned} \tag{1.17}$$

To provide evidence not merely of a failure in the bootstrap procedure, but of the exact failure identified above, I construct a synthetically corrected bootstrap estimator by calculating T_N^{t*}, R_N^{t*}, and $(\sigma_{1N}^t)^2$ directly in each bootstrapped sample. The synthetically corrected bootstrap estimator is given by

$$\hat{\tau}_s^{t*} = \hat{\tau}^t + T_N^{t*} + (\sigma_{1N}^t)^2 + R_N^{t*} \tag{1.18}$$

This estimator can be thought of as what a correct bootstrap procedure analogous to the proposed bootstrap would produce, if it were possible to correct the procedure without conditional mean estimation. All results that follow come from a 10,000-iteration Monte-Carlo simulation, with $\tau^t = 5$ and 200 bootstraps per iteration.

Figure 1.1 represents a baseline case, with $N = 1000$ and an equal number of treated and control units. The proposed bootstrap (histogram in red) consistently underestimates the true variance of the matching estimator by a significant margin. The synthetically corrected bootstrap (turquoise) correctly estimates the target variance. Given the theoretical underpinnings of the failure, this is expected. The missing power in $(\sigma_{1N}^t)'$ will result in underestimated variance whenever $\sum_{i=1}^{N}(D_i + (1 - D_i)\frac{1}{M}K_i) < \sum_{i=1}^{N}(D_I + (1 - D_i)\frac{1}{M}K_i)^2$, and this will almost always occur when the ratio of treated to control units is close to one.

This logic suggests that in settings with significantly more control observations (i.e. $\alpha << 1$), the proposed bootstrap will underestimate the target variance by a smaller

amount. This is because decreasing α causes the probability of any K_i exceeding 1 to decrease. Simulations confirm this conjecture. Figure 1.2 is the same simulation as Figure 1.1, except with 3 times as many control units for an α of $\frac{1}{3}$.

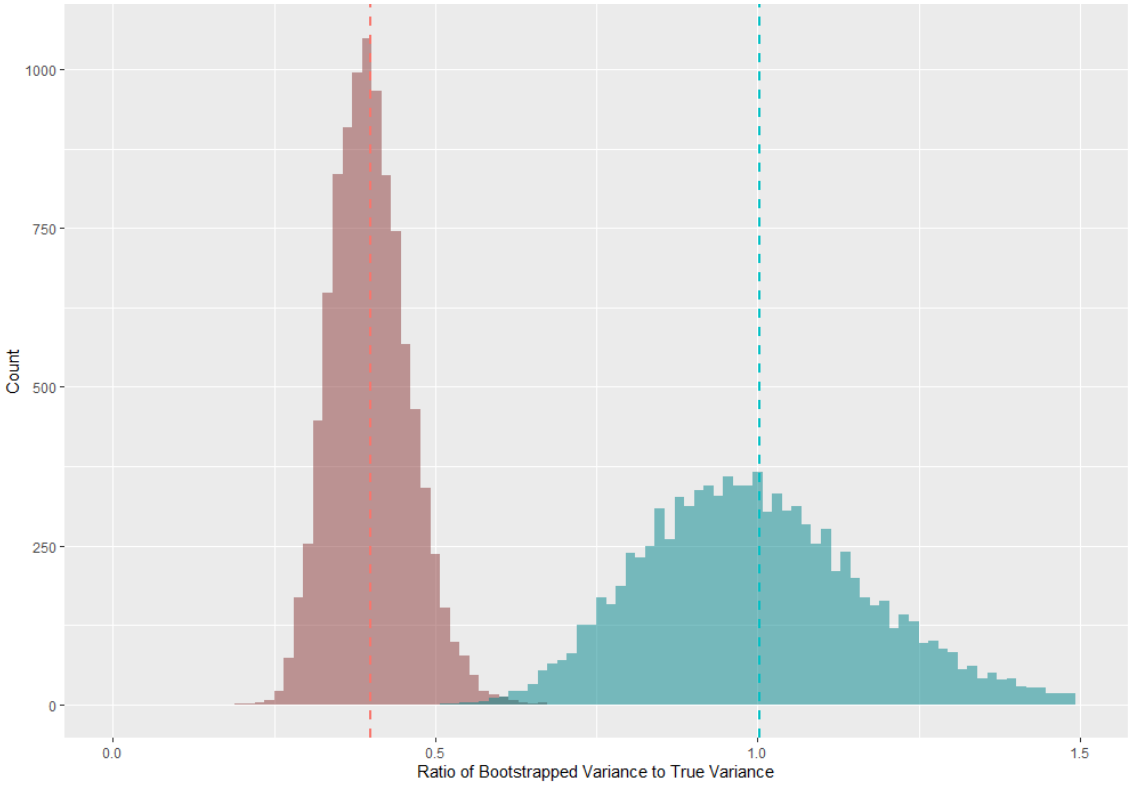

Figure 1.1 Proposed and Synthetically Corrected Bootstrap ($\alpha = 1$)

The rate at which the proposed bootstrap approaches the correct variance as α approaches 0 is very slow. Figure 1.3 displays the results from a simulation with $\alpha = 0.05$, and still the proposed bootstrap underestimates the target variance by a problematic amount.

In the appendix, I present simulation results for the estimation of the average treatment effect on the untreated population (the ATC). The Abadie and Imbens (2008) data generating process causes the failure in the proposed bootstrap to disappear due to the degenerate distribution of $Y(1)$. With a data generating process designed to be analogous to the Abadie and Imbens (2008) process but for ATC estimation, exactly the same failure would obtain.

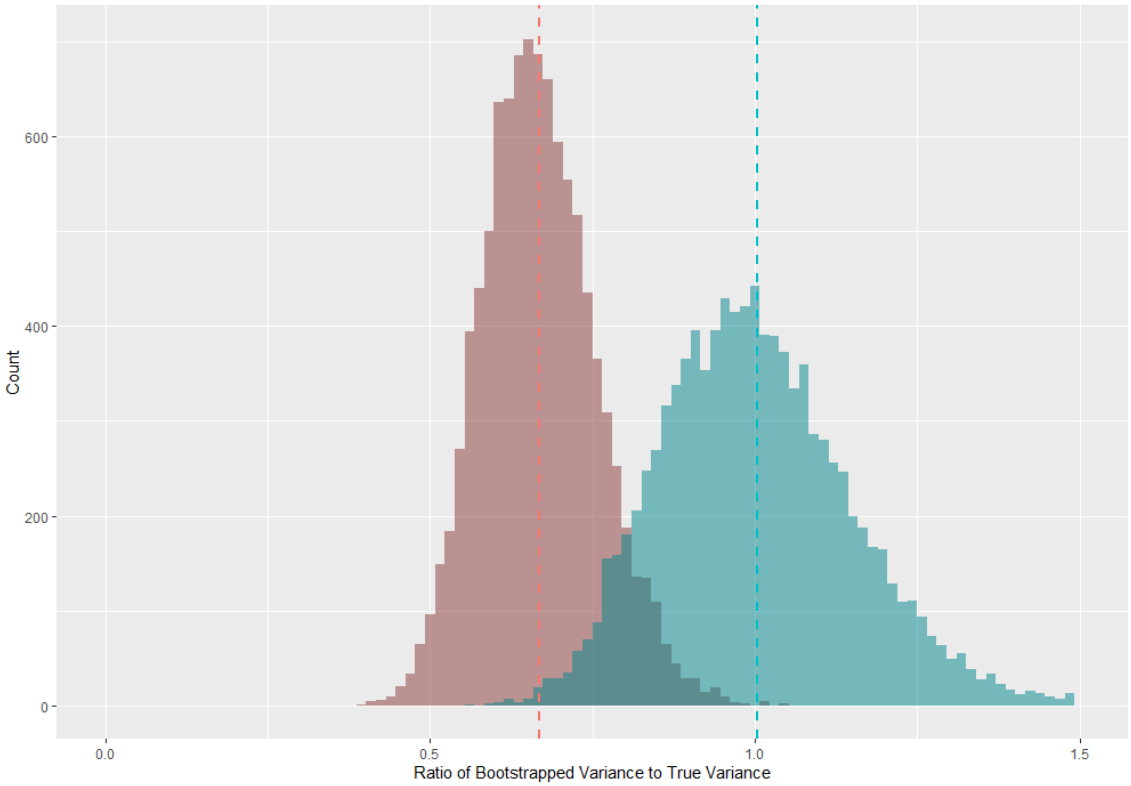

Figure 1.2 Proposed and Synthetically Corrected Bootstrap ($\alpha = 1/3$)

1.5 Conclusion

I proposed a *prima facie* reasonable bootstrap for estimating the variance of the matching estimator of the ATT when bias correction is unnecessary. I identified a flaw in the procedure, which is not a failure to estimate the distribution of K_i as in the naive bootstrap. Instead, the failure is related to the way in which control (treated) observations contribute variance to the final estimator of the ATT (ATC). The distribution of K_i is an important component of the variance of $\hat{\tau}^t$ because it determines how important the idiosyncratic error of unit i is to that variance. The proposed bootstrap fails to correctly perturb the idiosyncratic errors of control units, leading to a consistent underestimation of the true variance.

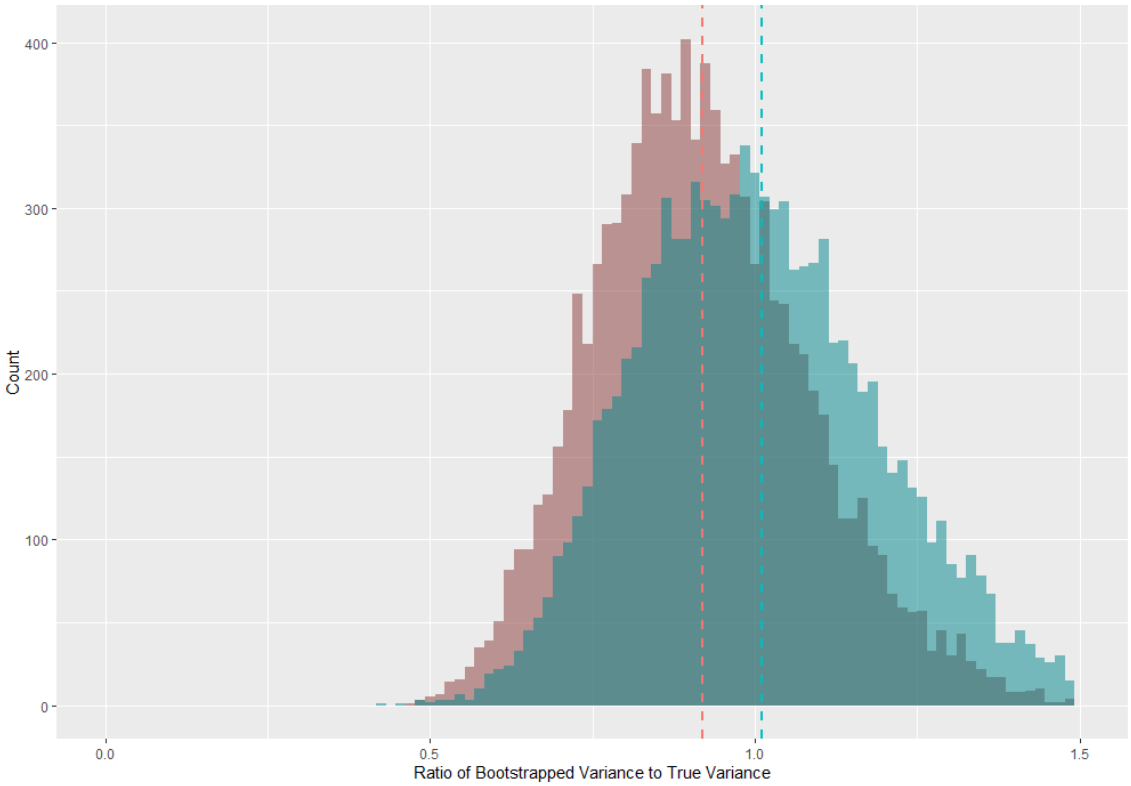

Figure 1.3 Proposed and Synthetically Corrected Bootstrap ($\alpha = 0.05$)

Otsu and Rai (2017) avoid this issue by sampling and bootstrapping two separate residuals. They construct these residuals by estimating the conditional mean functions $\mu(0, x)$ and $\mu(1, x)$. Avoiding the estimation errors associated with using the estimated functions $\hat{\mu}(0, x)$ and $\hat{\mu}(1, x)$ in the bootstrap was the was the leading source of potential improvements motivating this bootstrap. Thus, it appears that Otsu and Rai (2017) is currently the most efficient wild bootstrap procedure for estimating the variance of the matching estimator for treatment effects.

It is possible that a more complex bootstrap procedure may work specifically for estimating the average treatment effect for the whole population, when bias correction is unnecessary. Otsu and Rai (2017) requires only that $\hat{\mu}(d, x)$ satisfies a condition on convergence rates, which may be satisfied by the implicit estimator $\hat{\mu}(D_i, X_i) = Y_i\widehat{(D_i, X_i)}$ constructed in the normal matching estimator. If so, it would be possible to perform the

Otsu and Rai (2017) bootstrap without adding an extra step to estimate $\hat{\mu}(d, x)$ by using the implicit estimator. However, it is unlikely that performance improvements could be gained, as estimation errors would still affect the final bootstrap results and the estimation errors associated with the implicit estimator are likely to be large compared to explicit estimators of conditional mean functions. Otsu and Rai (2017) also note that when bias correction is unnecessary, it is possible to construct a valid subsampling procedure based on Politis and Romano (1994) that does not require estimation of $\hat{\mu}(d, x)$, although the computational cost of this procedure is significant compared to a wild bootstrap and the procedure is sensitive to the choice of subsample size when N is not large.

CHAPTER 2. MATCHING AS WEIGHT SELECTION: A FRAMEWORK FOR EVALUATING MATCHING ALGORITHMS

Due to non-smooth behavior of matching estimators, the bias/variance trade-offs associated with changes in the matching procedure are opaque. This leaves practitioners with limited guidance when choosing a matching procedure and its parameters. I cast matching estimators as a subset of a larger class of weighting estimators and use insights gained from considering optimal weights to offer further guidance in selection of matching procedures, selection of smoothing parameters, and potentially fruitful directions for future research.

2.1 Introduction

Matching estimation techniques have significant intuitive appeal and are relatively easy to implement. It is thus no surprise that despite Abadie and Imbens (2006) showing that they fail to reach the semi-parametric efficiency bound, they remain a popular approach to program evaluation. Perhaps due to the intuitive simplicity, a number of different approaches to matching have been proposed, none of which are obviously more or less plausible than others.

Practitioners today face a choice set that includes $M-$nearest neighbor matching (Abadie and Imbens, 2006), caliper matching (Cochran and Rubin, 1973), radius matching (Dehejia and Wahba, 1999), coarsened exact matching (Iacus et al., 2009), matching on the propensity score (Rosenbaum and Rubin, 1983, 1985) and genetic matching (Diamond and Sekhon, 2013). In addition, each of these procedures requires the choice of at least one smoothing parameter (e.g. number of matches, kernel and bandwidth, degree of coarsening). King

et al. (2011) suggests that researchers should conduct an "extensive, iterative, and typically manual search across different matching solutions," but this is unrealistically difficult to execute in practicer, and it is not clear what one should look for in this search.

In this chapter I aim to aid researchers facing this choice set by developing a framework that shrinks the relevant search space, identifying 'directions' within that space in which improvements are more or less likely to be found. This is accomplished by casting matching procedures as weight selectors which are followed by simple weighted difference-in-means estimation. Recasting matching estimators in this way allows me to identify infeasible optimal weights, and use the deviations from optimal weighting to generate insights about competing matching procedures.

This chapter's main contribution is to derive weights which are optimal in the sense of reducing mean-squared error (MSE), weights that are sometimes estimable[1]. First, I show that in the unconstrained case optimal weights are nonzero (outside of a degenerate case). I extend the result and prove that - subject to mild regularity conditions - the MSE-optimal weights are nonzero in situations that closely approximate those that apply to weights generated by matching. Using this insight, I develop a illustrative 'augmented' matching algorithm and verify through simulations that it behaves as my results suggest, confirming the validity of said insights. Further, the illustrative procedure sheds light on how important it is to avoid nonzero weights as features of the data-generating process change.

Overall, my results suggest that some form of kernel matching is likely to be most promising current approach in practice, as well as the most promising approach for further development of matching procedures. This is primarily due to the flexibility inherent in kernel matching, and echoes results from Armstrong and Kolesár (2018), who arrive at their conclusions from the consideration of worst-case MSE for weighting estimators in general.

[1]However, due to the form of the optimal weight functions, the cumulative effect of estimation errors is likely to limit the gains from using estimated optimal weights.

The approach I take necessitates conditioning on the sample, which has both pros and cons[2]. As Armstrong and Kolesár (2018) note, conditioning on the sample and realized treatment assignments takes into account the finite-sample possibility that imbalance may be present even with random assignment, but also precludes the use of the propensity score to gain efficiency. Like Armstrong and Kolesár, I do not intend to argue for or against conditioning on the sample - both approaches are valuable for understanding the behavior of program evaluation estimators.

This chapter contributes to a robust literature that offers advice to researchers choosing matching procedures and smoothing parameters. One strand of this literature contributes via simulation studies which contrast different matching procedures and smoothing parameters. For instance, Huber et al. (2013) generates data intended to replicate the features of a labor market dataset from Germany, and finds that radius matching with a regression adjustment performs best overall. Zhao (2004) considers the choice of the distance metric used in the matching procedure, a question that has received surprisingly little attention. King and Nielsen (2016) argues that propensity-score based pruning methods are inferior to other pruning methods (a claim presaged by Hahn, 1998).

Another strand of literature considers the use of weighting estimators for treatment effects more generally, without a strong focus on the connection to the method of matching. Most recently, Kallus (2016) and Armstrong and Kolesár (2018) develop methods of choosing weights that minimize worst-case MSE. Armstrong and Kolesár (2018) go on to provide asymptotically valid confidence intervals for a class 'minimax' optimal estimators they propose. Such 'minimax' estimators are designed to limit the MSE of an estimator subjected to a 'worst-case' data-generating process, characterized by a smoothness restriction on the conditional mean function. Hazlett (2016) develops a method to determine which weights achieve unbiased estimation of the average treatment effect on the treated. Hainmueller (2012) proposes a weight-selection algorithm that determines weights based on

[2]For a detailed discussion of this point, see Abadie et al. (2014))

moment conditions selected by the researcher, and Chan et al. (2015) employs a similar approach to develop a globally efficient calibration estimator. In contrast to this chapter, this strand of literature is generally concerned with cases where there is misspecification in either the regression function or the propensity score model.

The remainder of the chapter is organized as follows. Section 2.2 introduces notation and shows how matching estimators can be represented as weight selection procedures. In Section 2.3, I derive optimal weights for unconstrained and constrained cases, prove that optimal constrained weights are nonzero subject to mild regularity conditions, and offer some intuition for why this is true. In Section 2.4, I use that intuition to develop a simple 'augmented' matching procedure, and contrast it's behavior with nearest-neighbor matching and caliper matching through simulation. Finally, in Section 2.5 I conclude and suggest some directions for future research.

2.2 Setup & Notation

My notation closely follows Otsu and Rai (2017). We observe a dataset of size N, consisting of N_1 units that received treatment and N_0 units that did not. For each unit $i = 1, ..., N$, we observe a binary treatment indicator D_i, a covariate (potentially vector-valued) X_i, and an outcome,

$$Y_i = \begin{cases} Y_i(0) & \text{if } D_i = 0, \\ Y_i(1) & \text{if } D_i = 1 \end{cases}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for unit i if $D_i = 1$ and $D_i = 0$ respectively. Given this sample, we seek to estimate the average treatment effect for the treated population[3] (henceforth, the ATT)

$$\tau^t = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right]$$

[3]Extension of my results to the case of the average treatment effect for the untreated population (the ATC) is straightforward. Extension to the case of the average treatment effect for the whole population (the ATE) is less so, but can most easily be recovered by noting that the ATE is a weighted average of the ATT and the ATC.

To gain traction on this estimand, it is necessary that treatment is partially unconfounded[4], that the probability of treatment assignment is bounded away from 1, and that the sample is composed of conditionally independent draws from the population distribution. Formally,

A1. D is independent of $Y(0)$ conditional on $X = x$.

A2. $\Pr[D = 1 \mid X = x] < 1 - c$ for some $c > 0$.

A3. Conditional on $D_i = d$, the sample consists of independent draws from $Y, X \mid D = d$ for $d \in \{0, 1\}$.

These assumptions are standard in the matching literature. Assumptions A1 and A3 together are often referred to as 'selection on observables'. They ensures that the potential outcomes of two observations i and j will be equal if $X_i = X_j$, a requirement for matching estimators to be asymptotically consistent. Assumption A2, often called the 'overlap' condition, ensures that there are no portions of the covariate space in which all units are treated. If overlap does not hold, the ATT is not identified for subsets of the covariate space.

Before casting matching as a weight-selection procedure, let $\mu(x, d) = \mathbb{E}[Y \mid X = x, D = d]$, $\sigma^2(x, d) = \text{Var}(Y \mid X = x, D = d)$, and further let $\varepsilon_i = Y_i - \mu(X_i, D_i)$. Let $\mathbb{I}\{A\}$ be the indicator function that returns 1 when A is true, and 0 otherwise. Let $|x| = (x'x)^{1/2}$ be the standard Euclidean vector norm. Let $\mathcal{J}_M(i)$ be defined as

$$\mathcal{J}_M(i) = \left\{ j \in \{1, ..., N\} : D_j = 1 - D_i, \sum_{D_l = 1 - D_i} \mathbb{I}\{|X_l - X_i| \leq |X_j - X_i|\} \leq M \right\} \quad (2.1)$$

The standard M-nearest neighbor matching estimator for the ATT is then given by

$$\hat{\tau}^t = \frac{1}{N_1} \sum_{D_i = 1} \left(Y_i - \widehat{Y_i(0)} \right) \quad (2.2)$$

[4]Full unconfoundedness would be necessary when estimating the ATE.

where

$$\widehat{Y_i(0)} = \begin{cases} Y_i & \text{if } D_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } D_i = 1 \end{cases}$$

In simple terms, the M-nearest neighbor matching estimator imputes the value of $\widehat{Y_i(0)}$ as the average outcome values of the M control units with covariates closest to X_i. To recast matching as a weight-selection procedure, I will make use of an alternative representation from Abadie and Imbens (2006). First, define,

$$K_M(i) = \sum_{j=1}^{N} \mathbb{I}\{i \in \mathcal{J}_M(j)\} \tag{2.3}$$

$K_M(i)$ tracks the number of times that unit i is used as a match for another unit. To illustrate with a degenerate case, if $N_1 = 10$ and $N_0 = 1$, all 10 treated units would be matched to the single control unit. The value of $K_M(i)$ for that control unit would then be 10. Since it is more common to match with replacement, even in non-degenerate cases $K_M(i)$ can often be larger than 1. By convention, when estimating the ATT, $K_M(i) = 0$ when $D_i = 1$. It is straightforward to show that (2.2) can be represented as

$$\hat{\tau}^t = \frac{1}{N_1}\sum_{D_i=1} Y_i - \frac{1}{MN_1}\sum_{D_i=0} K_M(i)Y_i \tag{2.4}$$

By the nature of the M-nearest neighbor matching estimator, $\sum_{D_i=0} K_M(i) = MN_1$. Letting $k_i = K_M(i)/\sum_{D_i=0} K_M(i)$, we can rewrite (2.4) as

$$\hat{\tau}^t = \frac{1}{N_1}\sum_{D_i=1} Y_i - \sum_{D_i=0} k_i Y_i \tag{2.5}$$

which is a weighted difference-in-means estimator.

2.3 Optimal Weights

2.3.1 Unconstrained Weights

It is natural to ask at this point what the optimal value of the vector k_i is, and minimizing MSE is a natural objective to consider. Following Abadie and Imbens (2006), I characterize

the MSE of (2.5) in a useful way. Define the sample average treatment effect on the treated (SATT):

$$\overline{\tau^t(X)} = \frac{1}{N_1}\sum_{i=1}^{N}\left(\mu(1,X_i)-\mu(0,X_i)\right)$$

and note that the difference between $\hat{\tau}^t$ and the SATT is

$$\hat{\tau}^t - \overline{\tau^t(X)} = \frac{1}{N_1}\sum_{D_i=1} Y_i - \sum_{D_i=0} k_i Y_i - \frac{1}{N_1}\sum_{D_i=1}\mu(X_i,1) + \frac{1}{N_1}\sum_{D_i=1}\mu(X_i,0) \tag{2.6}$$

Recall that $\varepsilon_i = Y_i - \mu(X_i, D_i)$, and decompose the first term above to get

$$\hat{\tau}^t - \overline{\tau^t(X)} = \frac{1}{N_1}\sum_{D_i=1}\varepsilon_i - \sum_{D_i=0} k_i Y_i + \frac{1}{N_1}\sum_{D_i=1}\mu(X_i,0)$$

The error associated with $\hat{\tau}^t$ is thus

$$\hat{\tau}^t - \tau = \left(\overline{\tau^t(X)} - \tau\right) + \frac{1}{N_1}\sum_{D_i=1}\varepsilon_i - \sum_{D_i=0} k_i Y_i + \frac{1}{N_1}\sum_{D_i=1}\mu(X_i,0)$$

This offers a clear intuitive understanding of what an optimal vector of weights k_i would do. We cannot affect the value of $\left(\overline{\tau^t(X)} - \tau\right)$ through our estimation procedure - it is a function of the sampling procedure. The role of k_i is to turn $\sum_{D_i=0} k_i Y_i$ into an estimate of $\frac{1}{N_1}\sum_{D_i=1}\mu(X_i,0)$, the average of the unobserved counterfactual outcomes for the treated arm. The problem of minimizing MSE is isomorphic to the problem of setting the weighted sum of random variables to be as close as possible to some constant value. For ease of exposition, let $\frac{1}{N_1}\sum_{D_i=1}\mu(X_i,0) = \overline{\mu(X_{D_1},0)}$. Minimizing the MSE of the estimator in (2.5) is equivalent to solving

$$\min_{k_i} \mathbb{E}\left[\left(\overline{\mu(X_{D_1},0)} - \sum_{D_i=0} k_i Y_i\right)^2\right] \tag{2.7}$$

The solution to this problem leads to my first result:

Theorem 1 *Let $\sigma^2(X_i, D_i) = \sigma_i^2$. If $\sigma_i^2 \neq \sigma_j^2$ in general, the weights $\{k_i\}$ that solve the minimization problem in (2.7) are given by:*

$$k_i^* = \overline{\mu(X_{D_1},0)}\mu(X_i,0)\frac{\prod_{j\neq i}\sigma_j^2}{\sum_{i=1}^{N}\left(\mu(X_i,0)^2\prod_{j\neq i}\sigma_j^2\right) + \prod_{i=1}^{N}\sigma_i^2}$$

All proofs are relegated to the appendix. The requirement that $\sigma_i^2 \neq \sigma_j^2$ rules out the simplest form of homoskedasticity and is required to prevent the minimization problem from becoming degenerate. If σ_i^2 is a non-degenerate function of the covariate vector X_i, Theorem 1 holds and the k_i^* is at least in principle identified.

Some features of the optimal weight vector are worth noting at this point. As one would expect, if σ_i^2 is lower than σ_j^2, k_i^* will be larger than k_j^* if i and j have equivalent conditional means. In addition, the optimal weight vector $\{k_i^*\}$ contains no zero elements unless $\mu(X_i, 0) = 0$ for some i, or $\overline{\mu(X_{D_1}, 0)} = 0$, both of which are degenerate cases.

2.3.2 Constrained Weights

Unconstrained weights are of limited use for evaluation of matching procedures, because without constraints weights can be negative and can sum to something other than 1. Neither of these outcomes is a 'legal' outcome of any commonly used matching procedure.

Different matching procedures have different finite-sample constraints. For instance, if one uses M nearest-neighbor matching, conditional on the sample the only weights that can be generated are integer multiples of $\frac{1}{MN_1}$. By way of contrast, kernel matching is capable of producing weights that lie anywhere on the interval $[0, 1]$.

However, weights that are optimal subject to sample-specific constraints are unlikely to be useful in comparisons of different matching procedures - at best, they may shed light on the trade-offs involved in the choice of smoothing parameters. These trade-offs are less opaque, so I will focus on constraints that are shared across matching procedures - in particular, that individual weights are non-negative and that the weight vector sums to 1. One can think of these as the 'asymptotic' constraints that obtain on matching procedures - if the sample size is unknown, these are the constraints that obtain for all matching procedures. Further refinement of the constraints is impossible without knowledge of the sample size.

Theorem 2 provides a characterization of the MSE-optimal weights, subject to the constraint that weights are non-negative and sum to 1.

Theorem 2 *The weights* $\{k_i^*\}$ *that solve the minimization problem in (2.7), subject to* $k_i > 0 \ \forall i$ *and* $\sum_{i=1}^{N} k_i = 1$*, are given by:*

$$k_i^{c*} = \frac{1 - Y_1 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} + \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} + \overline{\mu(X_{D_1}, 0)} \left(Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} \right)}{\sigma_i^2 \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} + \left(\sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} \right) \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} \right) - \left(\sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} \right)^2 \right)}$$
$$+ \frac{\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2}}{Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}} \frac{\sigma_1^2 (Y_1 - Y_i)}{\sigma_i^2} \tag{2.8}$$

Unfortunately, the form of k_i^{c*} is not illuminating - it is, for instance, not at all clear what guarantees that k_i^{c*} is non-zero. In order to derive a strict positivity result for k_i^{c*}, I require an additional assumption:

A4. $\overline{\mu(X_{D_1}, 0)}$ *lies strictly between* $\min_i [\mu(X_i, 0) \mid D_i = 0]$ *and* $\max_i [\mu(X_i, 0 \mid D_i = 0)]$.

This assumption, while it appears quite restrictive, is rather general. It can be thought of as a finite-sample analog to the overlap condition A2. In most cases, if A4 is violated it is likely that AI.2 is violated as well. However, it is possible that A4 is violated without violating A2 in very small samples, or with extremely large values of σ_i^2. In either case, if A4 is violated, matching estimation will likely perform poorly with any matching procedure.

With this additional assumption, a strict positivity result for $\{k_i^{c*}\}$ can be proven,

Lemma 1 *Given assumptions A1 through A4, the MSE-optimal weight vector* $\{k_i^{c*}\}$ *contains no zero elements.*

I again relegate the full proof of Lemma 1 to the appendix. In simple terms, the proof works by showing that a vector containing a zero element can always be modified in a way that both removes the zero element and strictly reduces MSE - similar to how proofs related to Nash Equilibria search for profitable deviations.

The reason Lemma 1 is true relates to the shape of the function that describes the change in MSE when weight is 'shifted' from one observation to another. This function is

strictly concave and is always strictly increasing in a neighborhood around a zero 'shift', for appropriately selected observations. Intuitively, the 'change in MSE function' has this shape because the variance of the resulting estimator increases as a function of the *squared* weight on observations, while the bias increases as a function of the weight itself. The MSE-optimal weight vector never generates an unbiased estimate of the ATT - as one would expect, it achieves the minimal MSE by making 'profitable' trade-offs between bias and variance until no such trade-off remains.

2.4 Evaluating Matching Procedures

2.4.1 'Augmented' Matching

Lemma 1 is an interesting result from the perspective of one seeking to evaluate matching procedures. To the best of my knowledge, no commonly used matching procedure is designed to lower the chances of zero elements in the weight vector. Indeed, in some common settings many matching procedures will generate a wealth of zeros in the weight vector. However, the proof of Lemma 1 makes clear that the optimal weights, while nonzero, are nonetheless vanishingly small when the 'quality' of a match[5] is poor. If one is doing nearest-neighbor matching with $M = 1$, it is entirely possible that zero weights are closer to optimal than the lowest positive weight that can be assigned. Nonetheless, Lemma 1 suggests that we should consider more carefully the situations that can cause zero weights, and consider whether those weights should truly be zero.

A less obvious insight from Lemma 1 is that units with similar conditional means and similar variances should receive similar weights. It is this insight which guides the 'augmented' matching algorithm proposed below. The algorithm below is to estimate the ATT, but the extension to other estimands is immediate.

[5] In contrast to Chapter 1, I refer here to 'quality' in the sense of how far $\mu(X_i, 0)$ is from $\overline{\mu(X_{D_1}, 0)}$, not how far $\mu(X_i, D_i)$ is from some matched unit's $\mu(X_j, D_i)$.

'Augmented' Matching for the ATT

1. Perform standard nearest-neighbor matching with $M = 1$.
2. For each treated unit i, define a distance $r_i = |X_i - X_{m(i)}| + \delta$, where $m(i)$ returns the index of the matched control unit from step 1.
3. Search for control units whose covariates lie within a ball of radius r_i around X_i. If such control units exist, assign them as matches for unit i as well.
4. Use the resulting matches and weight vector to estimate τ^t.

This procedure encapsulates nearest-neighbor matching as a special case (if δ is set to zero, it is numerically identical to nearest-neighbor matching with $M = 1$). The idea is that, having already matched X_i to $X_{m(i)}$, it is likely that units some small δ further away are good enough matches to generate a variance decrease that outweighs the bias increase that comes from making a worse match.

To be clear, I am not advancing this procedure as the best that can be done given these insights. Rather, it is to illustrate that the insights derived are valid, and that this framework for evaluating matching procedures works. As I will show with simulations, this augmented matching procedure is too simple, but it serves to refine the insights derived so far.

2.4.2 Simulation Evidence

For my simulations, I designed a data-generating process that allows for a number of modifications that shed light on the relative performance of nearest-neighbor and augmented matching. The basic framework is quite simple - the outcome variable Y_i is constructed as $Y_i = X_i + D_i\tau(X_i) + \varepsilon_i$. I vary the distribution of X_i and ε_i across simulations. $\tau(X_i)$ is defined as $\tau(X_i) = \mathbb{I}\{X_i \geq 0\}\left(X_i + 2X_i^2 - 0.4X_i^3\right)$, to generate a conditional average treatment effect function with significant heterogeneity.

First, I draw X_i from a uniform distribution between 0 and 6. For each unit i, I also draw σ_i^2 from a uniform distribution between 2 and 5, and then draw ε_i from a mean-zero normal distribution with appropriate variance. I consider two matching procedures - nearest-neighbor matching with M neighbors and augmented matching as described above. I consider five different values of M and δ. Finally, I consider two cases for the sample - one with 500 units in each treatment arm, and one with 250 treated units and 750 control units. Table 2.1 presents the results, with 1000 simulations in each row.

Table 2.1 High Variance, Uniform X

N_1/N_0	Matching Procedure		ATT MSE	ATC MSE	ATE MSE
1	NN Matching	$M = 1$	0.094	0.099	0.077
	NN Matching	$M = 2$	0.076	0.078	0.067
	NN Matching	$M = 3$	0.069	0.072	0.065
	NN Matching	$M = 4$	0.066	0.069	0.062
	NN Matching	$M = 5$	0.065	0.066	0.061
1	Augmented Matching	$\delta = 0.25$	0.063	0.146	0.080
	Augmented Matching	$\delta = 0.50$	0.063	0.234	0.102
	Augmented Matching	$\delta = 0.75$	0.063	0.352	0.131
	Augmented Matching	$\delta = 1.00$	0.063	0.480	0.163
	Augmented Matching	$\delta = 1.25$	0.064	0.605	0.194
1/3	NN Matching	$M = 1$	0.142	0.114	0.100
	NN Matching	$M = 2$	0.108	0.096	0.089
	NN Matching	$M = 3$	0.095	0.088	0.083
	NN Matching	$M = 4$	0.090	0.085	0.081
	NN Matching	$M = 5$	0.086	0.083	0.080
1/3	Augmented Matching	$\delta = 0.25$	0.078	0.202	0.147
	Augmented Matching	$\delta = 0.50$	0.078	0.293	0.198
	Augmented Matching	$\delta = 0.75$	0.078	0.412	0.266
	Augmented Matching	$\delta = 1.00$	0.078	0.541	0.340
	Augmented Matching	$\delta = 1.25$	0.078	0.664	0.410

When estimating the ATT, the 'quality' of a match is simply $|X_i - X_{m(i)}|$, while for the ATC the 'quality' is $|\tau(X_i) - \tau(X_{m(i)})|$. With this data generating process, the latter grows dramatically faster than the former with the difference between X_i and $X_{m(i)}$. This explains the relatively poor performance of augmented matching in the ATC case.

The ATT case makes clear that the idea of augmented matching works in some settings. Note that even with $M = 5$, augmented matching outperforms nearest-neighbor matching with any tested value of δ. This illustrates the chief advantages of augmented matching over changing M - it allows for a different number of matches to be found for any given unit, and has a large 'sweet spot' for values of δ when match 'quality' is relatively flat.

In a second simulation, I change the distribution of σ_i^2 to a uniform distribution between 1 and 2, significantly restricting the potential size of idiosyncratic errors. Otherwise, the data-generating process was unchanged. Table 2.2 reports the results.

Lowering the size of idiosyncratic errors on observations would be expected to reduce the *relative* importance of variance in determining total MSE. Thus, one would expect augmented matching to perform more poorly relative to nearest neighbor matching in this simulation, and that is precisely what is observed.

Table 2.2 Low Variance, Uniform X

N_1/N_0	Matching Procedure		ATT MSE	ATC MSE	ATE MSE
1	NN Matching	$M = 1$	0.017	0.018	0.014
	NN Matching	$M = 2$	0.014	0.014	0.012
	NN Matching	$M = 3$	0.012	0.013	0.011
	NN Matching	$M = 4$	0.012	0.012	0.011
	NN Matching	$M = 5$	0.012	0.012	0.011
1	Augmented Matching	$\delta = 0.25$	0.016	0.097	0.033
	Augmented Matching	$\delta = 0.50$	0.016	0.186	0.055
	Augmented Matching	$\delta = 0.75$	0.016	0.304	0.084
	Augmented Matching	$\delta = 1.00$	0.016	0.433	0.117
	Augmented Matching	$\delta = 1.25$	0.016	0.558	0.148
1/3	NN Matching	$M = 1$	0.025	0.021	0.018
	NN Matching	$M = 2$	0.019	0.017	0.016
	NN Matching	$M = 3$	0.017	0.016	0.015
	NN Matching	$M = 4$	0.016	0.015	0.015
	NN Matching	$M = 5$	0.086	0.083	0.080
1/3	Augmented Matching	$\delta = 0.25$	0.018	0.137	0.08
	Augmented Matching	$\delta = 0.50$	0.018	0.225	0.133
	Augmented Matching	$\delta = 0.75$	0.018	0.340	0.200
	Augmented Matching	$\delta = 1.00$	0.018	0.466	0.270
	Augmented Matching	$\delta = 1.25$	0.078	0.664	0.410

One potential issue with the augmented matching algorithm is that δ is fixed for all units. In practice, if the distribution of observations within the covariate space diverges significantly from a uniform distribution, this may cause augmented matching to perform quite poorly. In particular, in areas of the covariate space where observations are sparse, augmented matching is likely to make a small number of additional matches, and those additional matches are likely to be poor quality.

To investigate this possibility, in Table 2.3 I change the data-generating process, drawing X_i from a $N(0, 2)$ distribution. This generates a large mass of units around 0, with significantly fewer units available as one moves away from 0. I return to the high-variance case in terms of idiosyncratic errors, drawing σ_i^2 from a $U[2, 5]$ distribution.

Table 2.3 High Variance, Normal X

N_1/N_0	Matching Procedure		ATT MSE	ATC MSE	ATE MSE
1	NN Matching	$M = 1$	0.088	0.088	0.068
	NN Matching	$M = 2$	0.067	0.070	0.058
	NN Matching	$M = 3$	0.063	0.065	0.057
	NN Matching	$M = 4$	0.062	0.063	0.056
	NN Matching	$M = 5$	0.059	0.062	0.056
1	Augmented Matching	$\delta = 0.25$	0.062	0.414	0.146
	Augmented Matching	$\delta = 0.50$	0.062	0.415	0.147
	Augmented Matching	$\delta = 0.75$	0.062	0.406	0.145
	Augmented Matching	$\delta = 1.00$	0.062	0.391	0.141
	Augmented Matching	$\delta = 1.25$	0.062	0.371	0.137
1/3	NN Matching	$M = 1$	0.123	0.114	0.098
	NN Matching	$M = 2$	0.097	0.094	0.084
	NN Matching	$M = 3$	0.086	0.087	0.079
	NN Matching	$M = 4$	0.081	0.083	0.077
	NN Matching	$M = 5$	0.078	0.081	0.075
1/3	Augmented Matching	$\delta = 0.25$	0.076	0.430	0.276
	Augmented Matching	$\delta = 0.50$	0.076	0.441	0.282
	Augmented Matching	$\delta = 0.75$	0.076	0.435	0.279
	Augmented Matching	$\delta = 1.00$	0.076	0.421	0.271
	Augmented Matching	$\delta = 1.25$	0.076	0.404	0.262

Somewhat surprisingly, the story is largely unchanged from the previous simulations. The relative comparison between nearest neighbor matching and augmented matching is similar to before, and nearest neighbor matching is not noticeably outperforming relative to when covariates were distributed uniformly.

For thoroughness, Table 2.4 reports results from a final simulation that draws σ_i^2 from a $U[1, 2]$ distribution.

Table 2.4 Low Variance, Normal X

N_1/N_0	Matching Procedure		ATT MSE	ATC MSE	ATE MSE
1	NN Matching	$M = 1$	0.016	0.017	0.013
	NN Matching	$M = 2$	0.012	0.015	0.011
	NN Matching	$M = 3$	0.011	0.014	0.011
	NN Matching	$M = 4$	0.011	0.015	0.011
	NN Matching	$M = 5$	0.011	0.015	0.011
1	Augmented Matching	$\delta = 0.25$	0.016	0.373	0.103
	Augmented Matching	$\delta = 0.50$	0.016	0.374	0.104
	Augmented Matching	$\delta = 0.75$	0.016	0.365	0.102
	Augmented Matching	$\delta = 1.00$	0.017	0.349	0.099
	Augmented Matching	$\delta = 1.25$	0.017	0.330	0.095
1/3	NN Matching	$M = 1$	0.022	0.023	0.019
	NN Matching	$M = 2$	0.017	0.020	0.017
	NN Matching	$M = 3$	0.015	0.019	0.016
	NN Matching	$M = 4$	0.014	0.019	0.016
	NN Matching	$M = 5$	0.014	0.019	0.016
1/3	Augmented Matching	$\delta = 0.25$	0.020	0.361	0.212
	Augmented Matching	$\delta = 0.50$	0.020	0.373	0.218
	Augmented Matching	$\delta = 0.75$	0.020	0.367	0.216
	Augmented Matching	$\delta = 1.00$	0.020	0.354	0.209
	Augmented Matching	$\delta = 1.25$	0.020	0.337	0.200

It is interesting to note that when X is distributed normally, augmented matching performs better in the ATC case as δ increases - a reversal of the behavior observed when X was distributed uniformly. However, it is clear that augmented matching in this form is not an appropriate technique for the estimation of the ATC, due to the relatively higher importance of bias in that case.

2.4.3 Discussion

A clear implication of the simulation results is that the correct δ is different when estimating the ATT and the ATC. Given the significant similarities between augmented matching, radius matching, and kernel matching, it is likely that this is true for the latter procedures as well. To the best of my knowledge, choosing the bandwidth separately for the ATT and ATC is not a common approach in practice. Since the ATE is a weighted average of the ATT and ATC, such a proposal is likely worth serious investigation, but it is beyond the scope of this investigation.

The simulations make clear that the insights derived from considering the MSE-minimizing weight vector are valid. In particular, practitioners should use economic intuition and knowledge of the empirical context (where possible) to weigh the relative importance of idiosyncratic errors and bias in the sample. In settings where bias is likely to be of low importance (for instance, if it is likely that the treatment effect is constant across X and the parameter of interest is the ATT), it is more likely that MSE can be reduced by matching to multiple units, or using procedures like radius and kernel matching which have strong similarities to the 'augmented' matching studied here. The same conclusion holds when there are many more observations in the 'donor pool' than in the pool of units to be matched (e.g. when estimating the ATT with many more control units than treated units).

2.5 Conclusion

Taking an approach similar to that of Armstrong and Kolesár (2018) and Kallus (2016), I derived unconstrained MSE-minimizing weights, and MSE-minimizing weights subject to constraints that approximate those implied by many common matching procedures. Subject to a mild condition on covariate balance, MSE-minimizing weights are nonzero, and units with similar conditional means receive similar weights.

I use an illustrative, and very simple, 'augmented' matching procedure that builds in behavior meant to generate weights that are closer to MSE-optimal, and contrast it with

nearest neighbor matching in a variety of settings. I find that the 'augmented' procedure compares favorably to nearest-neighbor matching when idiosyncratic errors are an important driver of MSE, while significantly under-performing when bias is relatively more important.

While the 'augmented' matching procedure itself is unsuitable for practical use in its current form, it confirms the insights I derive from the contrast between MSE-minimizing weights and weights from nearest neighbor matching. Practitioners can generate potentially significant reductions in MSE by carefully considering what they can reasonably determine about the data generating process, whether from economic intuition, knowledge of the empirical context, or knowledge of the data gathering process. In particular, my results suggest that defaulting to M nearest neighbor matching with $M = 1$ is likely to leave efficiency gains on the table in many common settings, but is a good conservative approach. This echoes Armstrong and Kolesár (2018), who note that when the data generating process is sufficiently 'bad', the minimax optimal estimator is nearest neighbor matching with one match.

Further development of the 'augmented' matching procedure may be worthwhile. In particular, the implication that the optimal δ differs when estimating the ATT and the ATC offers a potential route to cure the under-performance observed when bias is important, through a data-driven selection of δ. It is worth investigating whether this extends to radius and kernel matching procedures as well.

CHAPTER 3. THE EFFECT OF TEACHER GENDER ON STUDENTS OF DIFFERING ABILITY: EVIDENCE FROM A RANDOMIZED EXPERIMENT

Gender dynamics may play an important role in the determination of student outcomes in education. Exploiting random assignment of students to teachers in a field experiment, I study heterogeneity in the impact of teacher gender on the math and reading test scores for primary school students of differing ability. I find that assignment to a female teacher is generally positive for male students while having no significant effect for female students. In addition, I find very little heterogeneity in the effect of teacher gender on the ability axis, suggesting that average effect estimates do not mask significant heterogeneity. My results are consistent with differential teacher behavior based on gender stereotypes, and somewhat inconsistent with differential student behavior based on gender stereotypes.

3.1 Introduction

Achievement on school tests has important implications for students in both the short and the long run. In the short run, test scores serve as signals to students about their ability and induce students to choose different educational paths (Mechtenberg, 2009; Lavy, 2008; Lavy and Sand, 2018; Terrier, 2016). In the long run, these choices have major implications for lifetime earnings and health outcomes (Joensen and Nielsen, 2016; Autor and Wasserman, 2013; Krueger, 2017). Gender dynamics between students and teachers can play a significant role in determining student test score outcomes (Dee, 2005; Lavy, 2008; Antecol et al., 2015; Terrier, 2016).

To date, the study of gender dynamics in the classroom has mostly considered average effects, which can mask significant heterogeneity (Bitler et al., 2006). It is possible that the effect of teacher gender on students might depend significantly on student ability, which would have important implications for policy - particularly with regard to addressing inequality. For instance, male and female teachers may internalize different gender stereotypes and thus react differently to low- or high-performing male or female students (Williams and Ceci, 2015), or students may internalize different gender stereotypes and thus be more or less receptive to teaching from teachers of a particular gender (Ouazad and Page, 2012).

In this chapter, I address this question by studying how the effect of assignment to a female teacher changes with both the gender and ability of a student, using data from a field experiment conducted to evaluate the Teach for America (TFA) Program. I estimate the Conditional Average Treatment Effect (CATE) of assignment to a female teacher, conditioning on student gender and on pre-treatment test score as a proxy for ability. The CATE parameter is ideal for this study because it is a policy-relevant parameter that directly addresses the question of how student ability changes the effect of teacher gender on student outcomes. My estimates show how the effect of being assigned to a female teacher changes with both student gender and student ability.

Exploiting random assignment of students to teachers in the data allows me to deploy non-parametric techniques that require the strong assumption of unconfoundedness rather than imposing functional form restrictions. While the data is not representative of the U.S. primary school student population overall, it is representative of the most disadvantaged students and schools - a subset of particular importance to policymakers. Students in these schools are less likely to continue on to higher education, and thus more likely to face the challenges facing individuals without a college education in modern society[1].

[1]Men with less than a four-year college education have seen a dramatic reduction in real income over the last decade (Autor and Wasserman, 2013), are less likely to enter the labor force (Krueger, 2017), and face increased risk of poverty, physical health problems, and mental health problems. The prospects for women with less than a four-year college education are significantly worse than for women with more education, but are less grim than those for men.

I find very limited heterogeneity in the effect of teacher gender on students with different levels of prior achievement. For male students, assignment to a female teacher has a nearly uniform positive impact on math test scores. In reading, there is a small positive relationship between student ability and the effect of assignment to a female teacher. For female students, there is notably more heterogeneity in the effect of teacher gender. In math, there is a stronger positive relationship between ability and the effect of teacher gender than for male students, and some indication that the lowest-performing female students might be harmed by assignment to a female teacher. In reading, there is a non-monotonic relationship between student ability and the effect of teacher gender.

My results echo much of the previous economics literature in finding no significant *average* effect of teacher gender on students. Outside of the bottom of the pre-treatment test score distribution, the effect of assignment to a female teacher does not significantly differ with student gender. At the very bottom of that distribution, female students may benefit less than male students from assignment to female teachers in math. Notably, for all students, the effect of assignment to a female teacher is either positive or insignificant, which suggests that biases such as those found by Lavy (2008), Terrier (2016), or Cappelen et al. (2019) are not present in primary school.

The remainder of this chapter is organized as follows. Section 3.2 reviews related literature. Section 3.3 discusses the data, the institutional background, and the experiment itself. Section 3.4 briefly introduces the theoretical framework for the CATE estimator and sets out my estimation strategy. Section 3.5 presents the main results. Section 3.6 considers possible mechanisms and policy implications. Finally, Section 3.7 concludes.

3.2 Related Literature

In this chapter I contribute directly to the literature that studies student/teacher dynamics based on demographic features, and indirectly to a related strand of literature that considers the underlying mechanisms.

Reduced form estimates of the effect of demographic matching between students and teachers go back to Ehrenberg et al. (1995), who found that demographic matching had little impact on student learning, but a significant impact on teacher perceptions of students, using NELS:88[2] data. Dee (2004) used Project STAR data to investigate the effect of teacher race on students, finding a positive effect of same-race teachers on math and reading for students. Dee (2005) exploited a unique feature of the NELS:88 data to control for student fixed effects, again finding that student/teacher demographic dynamics had significant effects on teacher perceptions. Dee (2007), restricting attention to gender dynamics, found that assignment to a same-gender teacher significantly improved student test scores, teacher perceptions of the student, and student engagement.

Tertiary education has also received significant attention. Bettinger and Long (2005) and Hoffmann and Oreopoulos (2009) studied the effect of instructor gender on undergraduate students using administrative data from different universities[3]. Hoffmann and Oreopoulos (2009) found that assignment to a same-sex instructor boosted relative student performance and likelihood of course completion, but had little impact on upper-year course selection. Bettinger and Long (2005) found very mixed results - their primary conclusion is that the effect of instructor gender changes dramatically based on the subject in question. For instance, they found strong positive effects on female students in math and statistics, and a weak effect in economics. They also add to the growing number of studies that find negligible effects of instructor gender on male students.

Carrell et al. (2010), exploiting random assignment of students to teachers at the U.S. Air Force Academy, found limited impacts of instructor gender on male students, but significant positive impacts on female students in math and science. In contrast to Hoffmann and Oreopoulos (2009), Carrell et al. (2010) finds significant impacts for upper-year course selection. Fairlie et al. (2014), using administrative data from a community college, found

[2]The National Educational Longitudinal Study of 1988 consists of a representative sample of students that were in 8th grade in 1988.

[3]Bettinger and Long (2005) uses data on full-time undergraduate students in Ohio during 1998 and 1999. Hoffmann and Oreopoulos (2009) uses data on students at the University of Toronto.

similar effects for instructor race - in particular, assignment to an instructor from an underrepresented minority group shrinks the performance gap between white and minority students.

In postgraduate education, Neumark and Gardecki (1998) found that job placement outcomes for female graduate students in economics were not significantly impacted by the addition of female faculty members or having a female dissertation chair, while finding limited evidence for positive effects on graduation time and graduation likelihood. Hilmer and Hilmer (2007) studied top-30 economics doctoral programs between 1990 and 1994, and found that female students with male advisors were significantly more likely to accept a research-oriented first job, but found little effect on early career publication success.

In recent years, some large additional datasets have become available to researchers. Egalite et al. (2015) uses administrative from the Florida public school system to find small but significant effects of teacher race/ethnicity on students. Winters et al. (2013), also using Florida public school data, find that assignment to a female teacher positively impacts both male and female students in math, primarily between the 6th and 10th grade levels.

One implication of this is that teacher gender may not have an effect on student outcomes before middle school. However, there remains some uncertainty about when children begin to understand or internalize gender stereotypes. Ambady et al. (2001) suggests that it begins around 10 years of age, while Steele (2003) finds evidence suggesting that it begins as early as 7 years of age. Antecol et al. (2015), using the same data as I use here, finds that female teachers have a negative impact on female students in math, and no impact elsewhere. They offer suggestive evidence that the underlying mechanism is math anxiety among female teachers.

The mechanisms underlying student/teacher gender or race dynamics remain an area of ongoing research. One of the most commonly proposed theories is that teachers serve as role models for demographically similar students (Hess and L. Leal, 1997), potentially increasing student motivation and ambition (Maria Villegas et al., 2012), or reducing the

effect of stereotype threat[4] (Steele, 1997; Beilock et al., 2010). An alternative theory is that demographic dynamics affect teacher expectations of students, and that these expectations have material influence on relevant student outcomes. Prior research has found that teacher expectations are influenced by demographic matching (Ouazad and Page, 2012; Ouazad, 2014; Gershenson et al., 2016). The impact of teacher expectations on students appears to be largely uncontroversial, but Mechtenberg (2009) develops a model of cheap-talk grading that generates the same kind of achievement gaps observed empirically.

Finally, it could be that teachers are less likely to exhibit biases against demographically similar students, either directly through biased grading behavior (Terrier, 2016; Lavy, 2008; Lavy and Sand, 2018) or through moderated responses to student misbehavior (Downey and Pribesh, 2004; Holt and Gershenson, 2017).

Ouazad and Page (2012) offers suggestive evidence that the effect of teacher gender on students may depend on the students as well. In an experiment designed to elicit student beliefs about teacher biases, they found that male students correctly expected female teachers to be biased against them, while female students incorrectly expected male teachers to be biased in their own favor.

Pinning down the active mechanisms is a significant empirical challenge. The data necessary to distinguish between different mechanisms is difficult to acquire. For instance, determining whether teachers demonstrate biases in grading behavior requires access to both teacher grades and anonymous grades, as in Lavy (2008), Terrier (2016), or Lavy and Sand (2018). Carlana (2019) uses the Gender-Science Implicit Association Test to measure teacher biases directly, and finds that biased teachers increase the gender gap in math performance in their classes. Bassi et al. (2018), using video of teachers in Chilean schools, finds that teachers pay more attention to, and interact more favorably with, boys than with girls. This 'attention gap' is correlated with the gender gap in math scores in Chile.

[4]Stereotype threat posits that when an individual feels that they run the risk of confirming stereotypes about their social group, they become more anxious about their performance, and this may hinder their performance at a particular task.

3.3 Data

3.3.1 The National Evaluation of Teach for America

The data comes from the Mathematica Policy Research, Inc (MPR) National Evaluation of Teach for America (NETFA) Public Use File[5]. The NETFA was a field experiment conducted in elementary schools from six regions of the United States between 2001 and 2003. The full study consists of a pilot study, conducted in Baltimore during the 2001-2002 academic year, and a follow-up full-scale study conducted in Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta during the 2002-2003 academic year. In total, 17 schools containing 98 classes and 1,938 students took part in the experiment.

In each region, schools that had at least one TFA teacher and at least one non-TFA teacher assigned to teach a class in the same grade were considered 'eligible' for the experiment. From the pool of eligible school-grade combinations, MPR selected a random sample to form an experimental group that was representative of the schools where TFA teachers tended to teach at the time[6]. If a school-grade combination was selected for inclusion in the experiment, students entering that school and grade were randomly assigned to the teachers allocated to that school and grade. Throughout the experimental year, MPR performed roster checks to enforce original classroom assignments.

After the random assignment to classrooms, students in experimental classrooms took math and reading tests based on the last school grade they had completed, which I will refer to as pre-treatment tests. At the end of the school year, students again took math and reading tests based on the school grade they had just completed. For the vast majority of the students in the sample, the pre- and post-treatment tests were the grade-appropriate Iowa Test of Basic Skills (ITBS). A small group of students took their tests in Spanish - for these students, the test was the Logramos test. Both tests are published by the same organization (Riverside Publishing), but are normed relative to different groups.

[5] https://www.mathematica-mpr.com/-/media/publications/data-sets/2017/tfapublicuse.zip

[6] The Teach for America program has expanded significantly since the experiment. The sample is likely not representative of 'TFA schools' today.

The original purpose of the NETFA experiment was to evaluate the effectiveness of the Teach for America program. As a result, the sample is not representative of the U.S. school population - it is representative of the schools that usually participate in the TFA program. While this prevents my results from generalizing to the broader school population, the students served by these schools are a subset of the student population on which policymakers have focused in the past.

3.3.2 Sample Statistics

The NETFA data includes detailed information on student and teacher characteristics. For students, it includes class type (bilingual/monolingual), student demographic characteristics, class size, and math/reading scores both before and after treatment. For teachers, it includes demographic characteristics, type of teacher certification (nontraditional/traditional), and years of experience[7]. In addition to the baseline data, I construct a classroom-level indicator variable for the presence of at least one disruptive student[8].

The test score variables deserve some further discussion. The data does not contain traditional test scores. Instead, there are raw counts for number of correctly answered questions and number of questions attempted, and a battery of transformed scores. The transformed scores include standardized score, grade equivalent, national percentile rank, and normal curve equivalent scores. For my investigation, I use normal curve equivalent scores as both pre-treatment conditioning and post-treatment outcome variables. The primary reason for this choice is that normal curve equivalent scores have the same equal-interval property that a z-score does. This is critical for estimation techniques that average outcomes.

[7]Seven classrooms experienced teacher turnover during the experimental year. Following Antecol et al. (2015), I code the teacher as being the first teacher without missing data. In all but one case, this is equivalent to the longest-serving teacher.

[8]I use disciplinary data to proxy for this. Specifically, if a class contained at least one student who was suspended or expelled during the course of the school year, I code that classroom as having been disrupted. Some classes contained students that are not part of the research sample, so some classes may be incorrectly coded as not disrupted.

Normal curve equivalent (NCE) scores are defined $NCE = 50 + 21.063 * ss$ where ss is the standard z-score. The choice of 21.063 as the multiplier ensures that, if the underlying standard scores are normally distributed, then a percentile rank of 1, 50, or 99 corresponds to a normal curve equivalent score of 1, 50, or 99 respectively. Close to 50, normal curve equivalent scores change more slowly than percentile ranks, while close to 1 or 99, they change more rapidly[9].

Some students in the sample have raw scores of 99. These scores are invalid - the highest possible raw score in the sample is 44 in reading and 50 in math (Penner, 2016). Approximately 19 (21) percent of the initial math (reading) sample is lost due to students with missing or invalid data. This is a slightly larger loss than Antecol et al. (2015) because they retained invalid test scores in their main specification[10].

Table 3.1 reports summary statistics for the variables of interest. Note that the math estimation sample and the reading estimation sample are not identical. In general, this is because students who recorded an invalid test score in math or reading did not always record an invalid test score in both subjects. In the interests of dropping as little data as possible, I retain students with invalid test scores in the 'wrong' subject when estimating the CATE for math or reading outcomes.

Table 3.2 reports the results of tests for mean differences between the full sample and the two estimation samples. I find very similar results to Antecol et al. (2015) in these tests. Sample attrition appears to be largely at random.

While there are some significant differences in means between the full and estimation samples, most are quantitatively small. The only exceptions are in pre-treatment math and reading scores - and this is entirely due to the removal of invalid test scores[11].

[9] If the underlying test scores are normally distributed, a percentile rank between 89 and 95 will be transformed into a normal curve equivalent between 75.8 and 84.6. A percentile rank between 40 and 59 will be transformed into a normal curve equivalent between 44.7 and 54.8.

[10] In a supplementary specification, Antecol et al. (2015) removed the invalid scores and did not see a large change in their results.

[11] Invalid raw scores of 99 were coded as normal curve equivalent scores of 0. Thus, removal of invalid scores will mechanically drive mean pre-treatment test scores up. The full-sample mean pre-treatment scores after removing invalid scores are essentially identical to the estimation sample means.

Contrasting the estimation samples with only those students who have invalid test scores tells a somewhat different story. Black students are slightly more likely than average to have recorded an invalid math score, while being slightly less likely to record an invalid reading score. Hispanic students display the reverse pattern - they are slightly more likely to record an invalid reading score, and less likely to record an invalid math score. Finally, there is a statistically significant difference in the mean class size between the math estimation sample and the sample of students with invalid math scores. This is likely because larger classes have more chances to draw an invalid score, rather than there being a causal relationship between class size and invalid scores.

Since I will be estimating treatment effects conditional on pre-treatment test scores, it is worth looking at the distribution of those scores in the data. Figure 3.1 presents histograms of the pre-treatment math and reading scores across the relevant estimation samples. The red dashed line indicates the 90th quantile of the pre-treatment test score distribution for each sample.

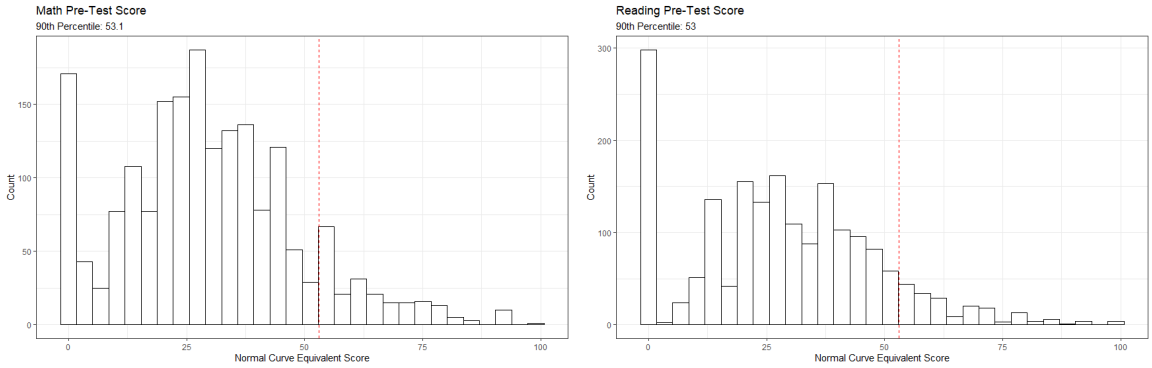

Figure 3.1 Pre-Treatment Test Score Distribution

Since I will be estimating treatment effects conditional on pre-treatment test scores, the relative lack of data in the upper half of the pre-treatment test score distribution has a direct impact on the variance of my estimates.

3.4 Estimation Strategy

Capturing the heterogeneity of a treatment effect has traditionally been done through the estimation of quantile treatment effects (QTEs), which describe the difference between quantiles of the outcome distribution for untreated and treated individuals. QTE estimation, however, allows for heterogeneity in the treatment effect across sub-populations that are not identifiable given covariates. For example, the QTE of assignment to a female teacher might be positive for students in the 60th quantile, negative for students in the 40th quantile, and zero for those in the 50th quantile - but it may not be possible to determine *a priori* whether a particular student was in any of those quantiles. In the context of my investigation, this is undesirable - I am interested in how the treatment effect of assignment to a female teacher changes with specific covariates (gender and pre-treatment test scores).

Thus, instead of the QTE, I estimate the Conditional Average Treatment Effect (CATE) function. The CATE is defined as the value of the Average Treatment Effect (ATE) within a sub-population defined by specific covariate values. While the CATE is not an entirely new parameter, often appearing as an intermediate estimand for ATE estimation (Heckman et al., 1997; Hahn, 1998), treatment of the CATE as a parameter of interest is relatively recent.

The chief difficulty in identifying the CATE is that unconfoundedness probably does not hold when conditioning on a strict subset of the available covariates. In the context of this investigation, it is unlikely that unconfoundedness holds conditional on only student gender and pre-treatment test scores. Abrevaya et al. (2015) provides a semi-parametric estimation procedure that accounts for this issue and allows for consistent estimation of the CATE parameter when conditioning on a subset of the covariates for which unconfoundedness does not hold.

I implement the Abrevaya et al. (2015) estimator and estimate the CATE of assignment to a female teacher. I condition on pre-treatment test scores after splitting the sample by student gender, recovering the CATE conditional on both covariates.

3.4.1 The Abrevaya et al. (2015) CATE Estimator

For compactness of notation, let Y_i be the post-treatment test score for student i, X_i be a vector of control covariates, and D_i be a binary indicator for the gender of student i's teacher ($D_i = 1$ if i was assigned to a female teacher, $D_i = 0$ otherwise). Let X_{1i} be a strict subset of X_i, containing only i's pre-treatment test score and an indicator for i's gender. Formally, the CATE is defined as

$$\tau(x_1) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_1 = x_1\right] \tag{3.1}$$

This parameter captures how the average treatment effect $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$ depends on the covariates contained in X_1 - in this context, how the effect of assignment to a female teacher changes with student gender and pre-treatment test scores. The Abrevaya et al. (2015) estimator of the CATE is

$$\hat{\tau}(x_1) = \frac{\frac{1}{nh^l}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\right)K_1\left(\frac{X_{1i}-x_1}{h_1}\right)}{\frac{1}{nh^l}\sum_{i=1}^{n}K_1\left(\frac{X_{1i}-x_1}{h1}\right)} \tag{3.2}$$

where $K_1(\cdot)$ and h_1 are respectively a kernel function and a bandwidth, l is the dimension of the vector X_1 (in this case, $l = 1$ because I condition on gender by splitting the sample, leaving only pre-treatment test score in X_1), and $\hat{p}(X_i)$ is an estimate of the propensity score[12]. Subject to mild regularity conditions on the first-stage propensity score estimation, Abrevaya et al. (2015) show that this estimator is asymptotically consistent for the CATE under the familiar unconfoundedness and sampling assumptions necessary for ATE estimation.

3.4.2 Identification Strategy

Intuitively, the identifying assumptions require that students who are assigned to a female teacher are comparable to students assigned to male teachers, conditional on pre-

[12] Abrevaya et al. (2015) considers both parametric and nonparametric estimation of the propensity score, and provides consistency results for both cases. While the nonparametric approach offers potential efficiency gains, it requires complicated transformations of discrete variables. In addition, it quickly runs into the curse of dimensionality when the set of covariates is of high dimension. As a result, I estimate the propensity score parametrically.

treatment test scores, student gender, and other covariates. If, for instance, students in one region had much stronger gender stereotypes and were also more likely to be assigned to a female teacher, unconfoundedness would likely fail. Without controlling for region effects in the estimation, the estimated effect of assignment to a female teacher would be biased downwards.

A major upside of the data used is that conditional on randomization block, students were assigned to teachers totally at random. This means that a number of potential confounders, like better students being assigned to female teachers[13], are not concerns. However, since the randomization is not unconditional, some potential sources of confounding remain. In particular, while students were randomly assigned to teachers, teachers were not randomly assigned to preparation pathways (i.e. TFA and non-TFA teachers are likely to be different), nor were they randomly assigned to schools or grades (i.e. it may be that teachers in one school are different from those in another school).

For TFA teachers, dealing with these issues is straightforward. TFA applicants in the experiment provided regional preferences, which allows for teachers to differ across regions, but not across schools within a region[14].

For non-TFA teachers, non-random assignment of teachers to schools or grades poses a more difficult problem. It is certainly possible that non-TFA teachers could select into different schools (or even grades) *within* a region, which would not be adequately controlled by a region indicator. However, it is hard to see why teachers would select differentially into schools within the population from which the sample was drawn. While teachers almost certainly select into or out of high-poverty schools, it is less clear that they select into different schools within the population of high-poverty schools - outside of simple geographic reasons, which are adequately controlled for by region indicators.

[13]Clotfelter et al. (2006) finds that male teachers are more likely to be assigned students with lower math and reading scores, so this would be a real concern with purely observational data.

[14]To be more specific, TFA applicants reported regional preferences as well as preferences for level of education (e.g. primary/middle/high school levels). Since the experiment considers only primary school students, the latter preferences cannot introduce confounding. I thank the TFA administrators for a thorough explanation of the application process at the time of the experiment.

This would seem to suggest that the propensity score should be estimated as a function of region indicators (and perhaps school/grade indicators). However, this goes too far towards treating the data as coming from a perfectly randomized experiment. Notably, some schools in the sample have no male teachers - using school indicators when estimating the propensity score would result in students from those schools having estimated propensity scores of either 0 or 1, which is far from credible. Even if there is differential selection of teachers into schools, it is very difficult to see how it could produce certain schools that would *never* have male teachers. The existence of schools with only female teachers is far more likely to be a result of the relative proportion of female primary school teachers in general, rather than evidence of a strong selection mechanism that eliminates male teachers entirely from some schools.

Additionally, for the purpose of estimating treatment effects, the goal of the propensity score estimation step is "to obtain estimates of the propensity score that balance the covariates between treated and control samples" (Imbens and Rubin, 2015). In finite samples[15] it is thus important to include not only covariates that potentially explain treatment assignment, but covariates that explain the outcome of interest - even if they are known not to play a role in treatment assignment. I thus estimate the propensity score with the following logistic regression:

$$\ln \frac{P(FTEACH_i = 1)}{1 - P(FTEACH_i = 1)} = \beta_0 + \beta_1 SC_i^{'} + \beta_2 TC_i^{'} + \beta_3 R_i^{'} + \beta_4 TFA_i + \beta_5 CS_i + u_i \quad (3.3)$$

where $FTEACH_i$ is an indicator for assignment to a female teacher, $SC^{'}$ is a vector of student covariates, $TC^{'}$ is a vector of teacher characteristics, $R^{'}$ is a vector of region dummy variables, TFA is an indicator for whether the teacher was a TFA teacher or not, and CS_i is the size of student i's class. Full details of this specification can be found in chapter A3, where I also consider some alternative specifications for the propensity score.

[15] With a sufficiently large sample, correctly specifying the propensity score model suffices to achieve covariate balance. However, in any finite sample, even one from a perfectly randomized experiment, there is no guarantee that weighting by the true propensity score will balance important covariates.

One potential issue facing any investigation that uses inverse probability weighting is the effect of very large or very small propensity scores. It is clear from equation (2) that if $\hat{p}(X_i)$ is very close to 0 (1) for treated (untreated) students, the importance of the outcomes for those students will be inflated significantly by the weighting procedure. Weights such as these lead to highly variable estimates, and may indicate a failure of the overlap condition. In the above specification, this is not a significant issue. To deal with the minority of students with extreme propensity scores, I set propensity scores above 0.95 (below 0.05) to 0.95 (0.05). The main specification is robust to different trimming behavior - in particular, dropping students with extreme propensity scores instead of changing their propensity scores does not have a noticeable effect on the results. One alternative specification, discussed in chapter A3, depends more strongly on trimming behavior.

3.4.3 Choice of Smoothing Parameters

The IPW-based estimator in (3.2) requires the choice of two smoothing parameters - the kernel and the bandwidth. Following Abrevaya et al. (2015), I set bandwidth to be a multiple of the sample standard deviation in the conditioning covariate (pre-treatment test score). In my main specification, the bandwidth is set to be half the sample standard deviation (approximately 9 for male students in math, for example). I use a Gaussian kernel:

$$K_g(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3.4}$$

In chapter A3, I report results for different bandwidths and kernels. As is often the case with kernel-based local averaging, bandwidth choice strongly influences the resulting estimates, while kernel choice generally does not have a strong effect. Smaller bandwidths produce more variable CATE estimates, which are often non-monotonic and can have extreme ranges. Larger bandwidths produce flatter CATE estimates, and mechanically force the estimated CATE function towards monotonicity. As bandwidth increases, the CATE estimator quickly becomes uninformative as to heterogeneity, essentially recovering an estimate of the ATE.

While overfitting is a valid concern, my main goal is not to provide another estimate of the average effect of teacher gender. Heterogeneity in that effect is my primary concern, and I thus err on the side of choosing a bandwidth that is too small for my main specification.

3.5 Results

3.5.1 Conditioning on Pre-Treatment Test Score

Figure 3.2 depicts the estimated CATE function for female students. Post-treatment math test scores are the outcome of interest, and the conditioning covariate is the student's pre-treatment normal curve equivalent test score in math. Pointwise valid confidence intervals are constructed using the asymptotic approximations from Abrevaya et al. (2015)[16]. As one would expect, given the distribution of pre-treatment test scores in the sample (Fig-

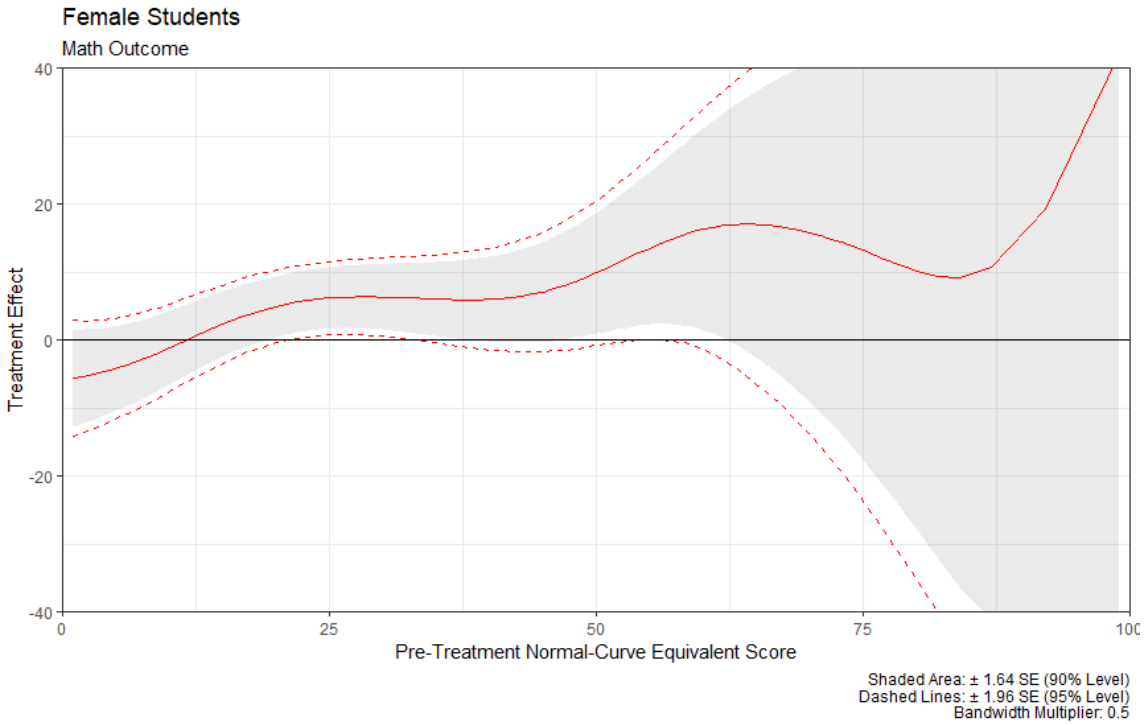

Figure 3.2 CATE (Math) for female students

[16]To the best of my knowledge, construction of uniformly valid confidence intervals for the Abrevaya et al. (2015) estimator is an open problem.

ure 3.1), the size of the confidence intervals grows rapidly once the pre-test score exceeds approximately 50, due to lack of data. Notably, the confidence interval for a pre-test score of 1 is relatively small, despite being a boundary point. This is largely due to the significant mass of students scoring 1 on the pre-test (also seen in Figure 3.1).

For the majority of students in this sample, I cannot reject the hypothesis that the true effect of being assigned a female teacher is zero. Indeed, while the confidence intervals here are pointwise valid, it is likely that uniformly valid confidence bands would be wider, and might not reject the hypothesis that the true effect of assignment to a female teacher is a *constant* zero across the pre-treatment test score distribution.

Qualitatively, while the majority of the point estimates are insignificant, the confidence intervals themselves suggest that if the true effect is not zero, female students at the very bottom of the ability distribution in math see less benefit from assignment to a female teacher than female students of higher ability. Outside of the very bottom of the ability distribution, there does not appear to be much, if any, heterogeneity in the effect of teacher gender on math test scores for female students. My results are reasonably consistent with the true CATE having a monotonic relationship between pre-test scores and the treatment effect. Indeed, particularly for TFA teachers, a possible conjecture is that students with higher ability are easier to teach effectively[17].

The implied average treatment effect[18] is around 0.25 standard deviations, or 4.5 points on the normal curve equivalent scale. While this is quite high, especially in comparison to Antecol et al. (2015), note that formally assessing the statistical significance of the implied ATE remains an open question. In light of the confidence intervals and the size of the implied ATE, it seems unlikely that the implied ATE would be statistically significant[19]. Restraining

[17]Since TFA is a *highly* selective program and primarily accepts the highest-achieving applicants, it is likely that those applicants were high-achievement students in primary school as well. Since they receive a relatively small amount of accelerated training in teaching, they may have an easier time understanding the difficulties faced by high-achieving students in their classrooms while struggling to understand those difficulties faced by the lowest ability students.

[18]The implied ATE is calculating by taking a weighted average of the CATE point estimates, where the weight on $\hat{\tau}(x_1)$ is equal the proportion of the sample with $X_1 = x_1$. It is the point estimate of the average treatment effect we would expect to see if the CATE point estimates are correct.

[19]I performed a standard non-parametric bootstrap for the implied ATE, and subject to the caveat that

the calculation to consider only point estimates below 55, thus excluding potentially extreme point estimates driven by lack of data, the implied ATE decreases to around 0.19 standard deviations (3.4 on the normal curve scale).

Figure 3.3 depicts the estimated CATE function for male students, again with math scores as the outcome of interest and conditioning covariate. The increase in the size of the confidence intervals starts earlier than in Figure 3.2, primarily because the male pre-test score distribution is skewed to the left relative to the full sample, which is in line with male students generally performing worse than female students in school. In addition, since no male in the sample scored higher than 92 on the pre-test, CATE estimates for pre-treatment test scores above 92 cannot be constructed.

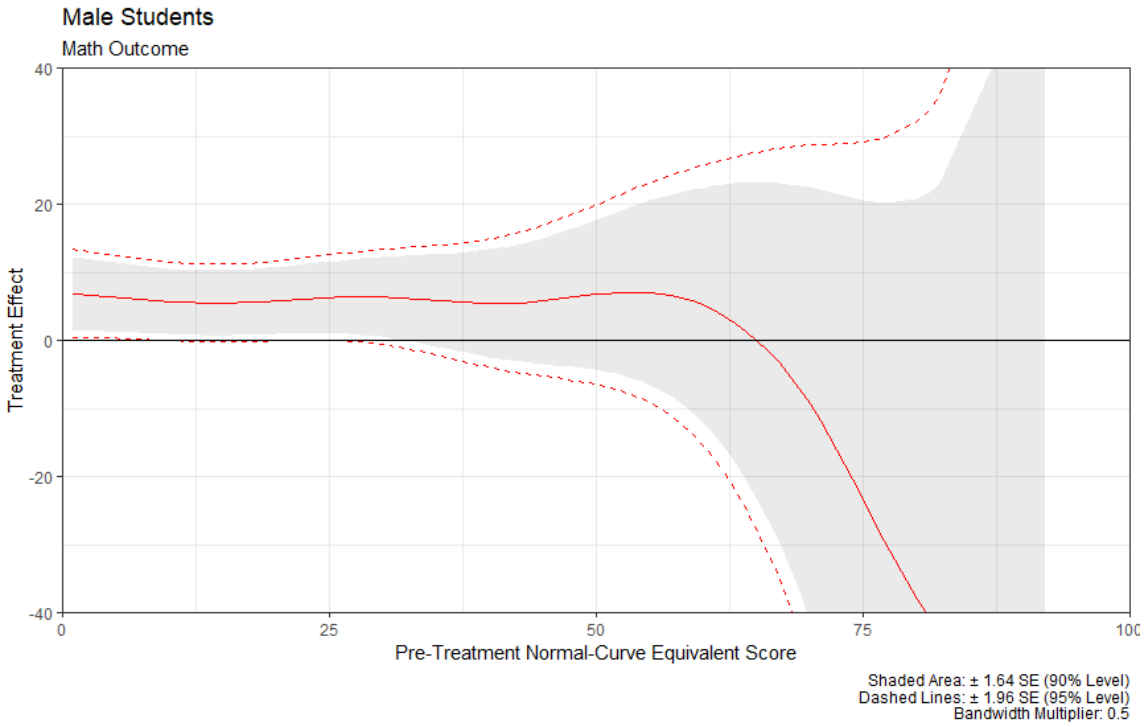

Figure 3.3 CATE (Math) for male students

In contrast to Figure 3.2, for the majority of the students in this sample the effect of assignment to a female teacher is at least marginally significant and positive. This is in

such a procedure is not currently known to be valid, the bootstrap results support this claim.

stark contrast to what one would expect if female teachers were biased against low-ability male students. If anything, my results so far would be consistent with a bias in the opposite direction - against low-performing or low-ability *female* students.

The implied ATE is approximately 0.25 standard deviations (4.7 on the normal curve scale). Considering only pre-test scores below 55 raises the implied ATE significantly to 0.33 standard deviations (6.0 on the normal curve scale). As before, it seems unlikely that the implied ATE would be statistically significant. Using the same rough rule of thumb that uniformly valid confidence bands would be larger, it is also unlikely that I would be able to reject the hypothesis that the true effect was a constant zero.

It is notable that, discounting the extreme point estimates arising from lack of data at the very top of the pre-treatment test score distribution, there is essentially no evidence of heterogeneity in the effect of teacher gender on male students. A male who scored 1 on the pre-test has nearly the same estimated CATE as one who recorded a score between 2 and 55. The only change is an increase in the size of the confidence intervals, which may be entirely due to the decrease in available data as test scores increase. The size of the positive effect is roughly the same as for female students in the middle of the pre-treatment test score distribution.

Figures 3.4 and 3.5 depict the estimated CATE functions for female and male students, respectively, with reading test scores as the outcome of interest and conditioning covariate. The first-stage propensity score model is the same as before except for the change from math to reading test score variables. For female students, there is noticeably more heterogeneity in the estimated CATE function, and it is no longer consistent with a monotonic relationship between treatment effects and pre-treatment test scores. The implied ATE is around 0.09 standard deviations (1.7 on the normal curve scale). A much smaller effect on reading than in math is consistent with previous literature studying the effect of teacher gender. Restricting attention to pre-test scores below 55 has almost no impact on the implied ATE. In contrast to previous literature suggesting that effects on

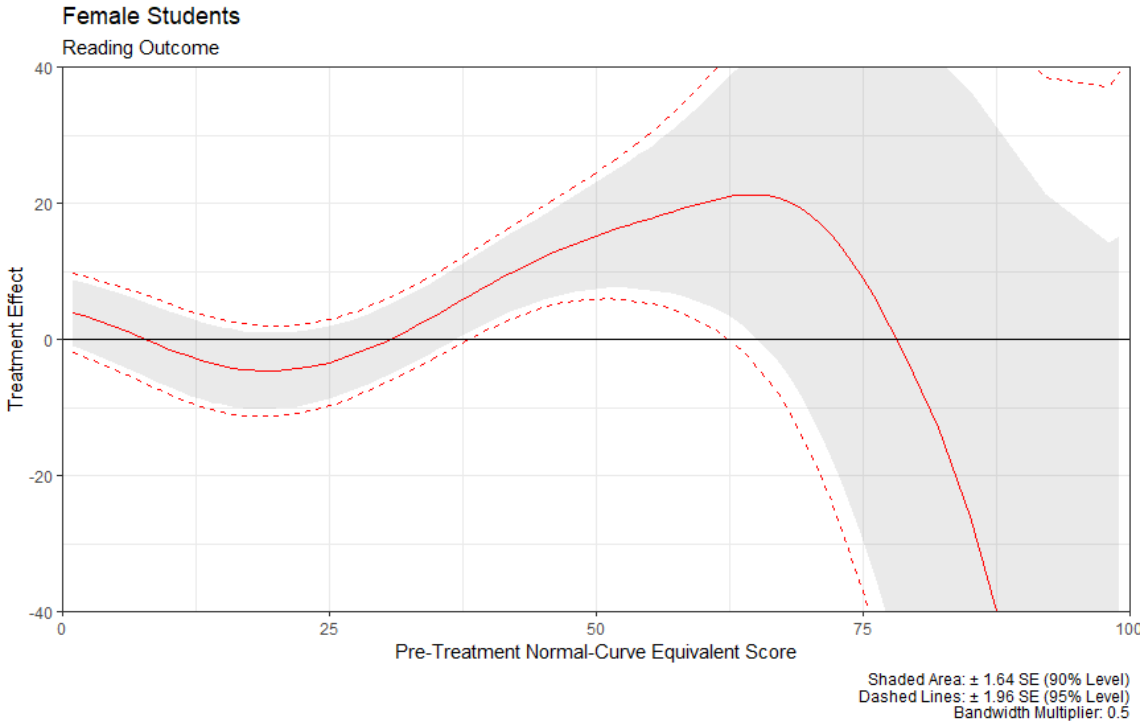

Figure 3.4 CATE (Reading) for female students

reading are non-existent, I find that female students with pre-treatment test scores in the middle of the distribution see a significant and large positive treatment effect.

For male students, the story appears largely the same as before. There is limited heterogeneity (although potentially more than in math). The estimated CATE is positive for almost all pre-treatment test scores below 55, as before, and the change in the CATE within that range is limited. As was the case with math results, the implied ATE for male students is relatively large - approximately 0.31 standard deviations (5.5 on the normal curve scale) for the full sample, and around 0.29 standard deviations (5.2) for students scoring less than 55 on the pre-test. Again, it is unlikely that the implied ATE is statistically significant.

3.5.2 Conditioning on Class Rank

To this point, I have been agnostic as to what might drive heterogeneity in the effect of teacher gender. Most of the standard mechanisms for teacher gender effects could plausibly

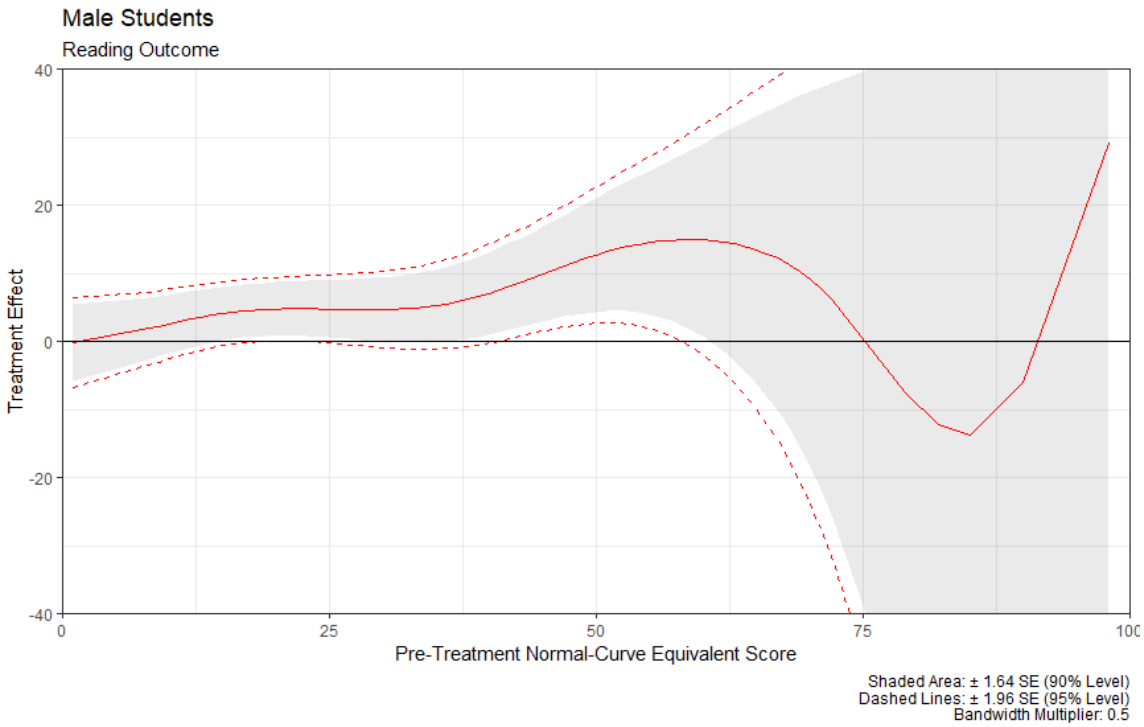

Figure 3.5 CATE (Reading) for male students

include heterogeneity on ability. Role model effects, for instance, might be stronger for high-ability students, or stereotype threat effects on women in math may be more powerful at the low end of the ability distribution. However, it is also possible that teacher behavior differs for students of different *perceived* ability - e.g. teachers may invest different amounts of effort in students they perceive as struggling or excelling.

Perceived ability may not closely track 'objective' ability as measured by pre-treatment test scores, or it may be that teachers care more about the ability of a student relative to the rest of the class, rather than relative to a national norm group. To investigate this possibility, I estimate the CATE functions as before, but replace the pre-treatment test score with a class rank variable constructed from the data[20]. Figure 3.6 presents the estimated CATE functions conditional on class rank for the four subsamples.

[20]Unfortunately, since some classes contain students not in the research sample, the accuracy of this variable is likely imperfect. If there is a correlation between student ability and whether a student was in the research sample, identification of the CATE may fail for this specification.

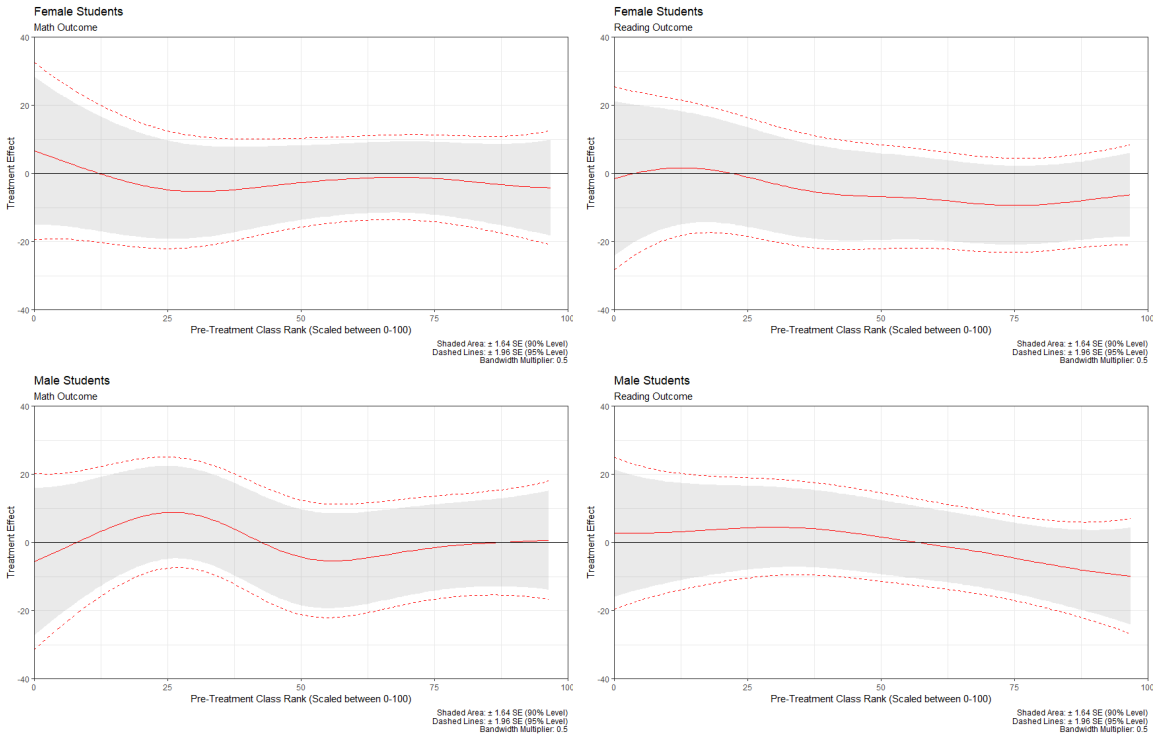

Figure 3.6 Conditioning on Class Rank

The class rank variable is scaled into a 'percentile' rank, with 0 being the worst student in the class and higher values reflecting higher within-class rankings, so the interpretation of the graphs is similar to before - and the results suggest that within-class performance is not correlated with the size of the teacher gender effect. Even with point-wise valid confidence bands, the hypothesis that the true effect conditional on class rank is a constant zero cannot be rejected in any sub-sample at the 95% level.

3.6 Discussion

Somewhat surprisingly, the overriding takeaway from this investigation is that there is very little heterogeneity in the effect of teacher gender on students of different levels of ability. Assignment to a female teacher is either neutral or positive for all students, and the

heterogeneity is largely confined to the different effects for male and female students. In math, male students see a uniformly positive effect from assignment to a female teacher, as do female students outside of the very bottom of the pre-treatment test score distribution. In reading, I find that students of either gender with pre-treatment test scores that are average compared to the national norm see positive effects from assignment to a female teacher, and the remainder of students see no significant effect.

The presence of significant effects on reading is surprising in light of the existing literature. It may be that, for relatively well prepared students, female teachers are more effective in teaching reading because they have internalized stereotypes labeling reading as an area where women are better. It may also be the students who have internalized such a stereotype, and exert more effort or are more engaged in reading when taught by a woman.

Differential teacher behavior could also explain why I find a positive effect on male students in math, but no significant effect for female students. Female teachers who view math as a 'male' subject might view low achievement from a male student as a sign that help is needed, while viewing low achievement in math from a female student as being expected. Unlike with the reading effects, it is difficult to see how traditional gender stereotypes about math might drive male students to be more engaged when taught by women.

In terms of policy implications, the most important implication is that male students benefit from assignment to female teachers, while female students appear largely unaffected. Primary school teaching is already an occupation dominated by women, and my results suggest that, if anything, this has benefited male students.

Since classes are generally not split by gender, consideration of teacher gender when assigning teachers to classes is unlikely to generate benefits overall. That said, Clotfelter et al. (2006) finds that male teachers are more likely to be assigned to classes with lower average math and reading scores. This kind of sorting is likely to have a negative overall effect on student achievement - while the very worst-performing female students might benefit from assignment to a male teacher, my results suggest that male students will

be harmed, and female students with higher scores may also be harmed relative to being assigned a female teacher. If anything, my results suggest that, all else equal, women should be preferred when seeking a teacher for a classroom of low-achieving students.

In terms of average effects, my results differ from those of Antecol et al. (2015), who find a negative association between assignment to a female teacher and a female student's test scores in math. Partially, this is due to consideration of different parameters. Antecol et al. (2015) consider estimates of what can be thought of as the effect of being a female student, and how that changes with teacher gender. In their specification, the estimated effect of being assigned to a female teacher is insignificant at conventional levels for all students, which is at least somewhat consistent with my results. More generally, the relative treatment effects for male and female students display the same relationship - males benefit more (or are harmed less) by assignment to a female teacher. Antecol et al. (2015) also provide suggestive evidence that the mechanism underlying their results is powered by stereotype threat, which falls in line with the hypothesis of differential teacher behavior proposed above.

As my sample is not representative of the U.S. student and teacher populations, it is possible that my results are driven by the difference between the population of disadvantaged schools and the broader U.S. school population. It is plausible, for instance, that teachers working in the most disadvantaged schools are less likely to be biased against (or more aware of their potential biases against) low-ability students. They may receive specialized training to help them effectively teach low-ability students that a teacher in a less disadvantaged school would not receive. The level of schooling may also play a role, as my sample consists entirely of primary school students between first and fifth grade. This may be too early for gender stereotypes to strongly affect gender dynamics between students and teachers, although Antecol et al. (2015) suggests otherwise. Different levels of schooling, and a sample more representative of the U.S. school population overall, provide exciting avenues to extend this research.

3.7 Conclusion

I estimate the Conditional Average Treatment Effect of assignment to a female teacher on students of different abilities, using data from the National Evaluation of Teach for America, a field experiment run between 2001 and 2003. I find little evidence of heterogeneity across students of different abilities, and a small degree of heterogeneity across students of different genders. Male students see a uniformly positive but marginally significant effect from being assigned to a female teacher in math, while female students see effects that are generally insignificant. In reading, students that are average relative to the national norm group see positive and significant effects from assignment to a female teacher, while the remainder of students see insignificant effects.

Overall, my results suggest that teacher gender effects in math do not significantly change with student ability, with what little heterogeneity there is being primarily on the gender axis. In reading, there is some evidence of heterogeneity along the ability axis, but much less difference between students of different genders. My results are most consistent with teachers internalizing traditional gender stereotypes regarding math and reading, and not at all consistent with the bias found in Cappelen et al. (2019).

Table 3.1 Descriptive Statistics

	Definition	n=1938 Full Sample	n=1596 Math Sample	n=1551 Reading Sample
Student Characteristics				
Female	1 if student is female, 0 otherwise	0.49 (0.50)	0.49 (0.50)	0.50 (0.50)
Black	1 if student is non-Hispanic black, 0 otherwise	0.67 (0.47)	0.66 (0.48)	0.70 (0.46)
Hispanic	1 if student is Hispanic, 0 otherwise	0.26 (0.44)	0.28 (0.45)	0.24 (0.43)
Class Size	Number of students in the classroom at the end of the experiment	25.1 (5.6)	24.9 (5.5)	25.2 (5.6)
Pre-Treatment Math	Normal Curve Equivalent (NCE) score on math pre-test	29.7 (18.6)	31.2 (18.2)	29.4 (17.4)
Pre-Treatment Reading	Normal Curve Equivalent (NCE) score on reading pre-test	28.8 (19.3)	29.5 (19.4)	29.9 (18.4)
Disrupted Class	1 if student was in the same class as another student who was suspended or expelled	0.45 (0.50)	0.46 (0.50)	0.47 (0.50)
Teacher Characteristics				
Female	1 if teacher is female, 0 otherwise	0.76 (0.43)	0.77 (0.42)	0.76 (0.43)
Black	1 if teacher is non-Hispanic black, 0 otherwise	0.50 (0.50)	0.48 (0.50)	0.51 (0.50)
Hispanic	1 if teacher is Hispanic, 0 otherwise	0.09 (0.29)	0.10 (0.31)	0.08 (0.28)
TFA	1 if the teacher is a TFA teacher, 0 otherwise	0.44 (0.50)	0.43 (0.50)	0.44 (0.50)
Certification	1 if the teacher has a traditional teaching certification, 0 otherwise	0.53 (0.50)	0.56 (0.50)	0.53 (0.50)
Experience	Years of teaching experience	6.42 (8.5)	6.2 (8.0)	6.19 (8.0)

Standard errors in parentheses.

Table 3.2 Mean Differences between Full and Estimation Samples

	Full vs Math Estimation	Full vs Reading Estimation
Student Characteristics		
Female	-0.003	-0.007
Black	0.015	-0.026^{*}
Hispanic	-0.023^{*}	$0.021^{\dagger}$
Class Size	0.233	$-0.096^{\dagger}$
Pre-Treatment Math	-1.579^{*}	0.248
Pre-Treatment Reading	−0.720	-1.122^{*}
Disrupted Class	−0.014	$-0.024^{\dagger}$
Teacher Characteristics		
Female	−0.005	0.005
Black	0.018	−0.008
Hispanic	−0.001	0.010
TFA	0.006	−0.007
Certification	$-0.024^{\dagger}$	0.009
Experience	-0.024^{*}	0.238

* denotes significance at the 5% level

† denotes significance at the 10% level

BIBLIOGRAPHY

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2009). A Martingale Representation for Matching Estimators. IZA Discussion Papers 4073, Institute of Labor Economics (IZA).

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614.

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505.

Adusumilli, K. (2017). Bootstrap inference for propensity score matching. Working paper.

Ambady, N., Shih, M., Kim, A., and Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science*, 12(5):385–390. PMID: 11554671.

Antecol, H., Eren, O., and Ozbeklik, S. (2015). The Effect of Teacher Gender on Student Achievement in Primary School. *Journal of Labor Economics*, 33(1):63–89.

Armstrong, T. B. and Kolesár, M. (2018). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness.

Autor, D. and Wasserman, M. (2013). Wayward sons: The emerging gender gap in labor markets and education. Technical report, Third Way.

Bassi, M., Mateo Díaz, M., Blumberg, R. L., and Reynoso, A. (2018). Failing to notice? uneven teachers' attention to boys and girls in the classroom. *IZA Journal of Labor Economics*, 7(1):9.

Beilock, S. L., Gunderson, E. A., Ramirez, G., and Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, 107(5):1860–1863.

Bettinger, E. P. and Long, B. T. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *The American Economic Review*, 95(2):152–157.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012.

Cappelen, A. W., Falch, R., and Tungodden, B. (2019). The boy crisis: Experimental evidence on the acceptance of males falling behind. Discussion Paper Series in Economics 6/2019, Norwegian School of Economics, Department of Economics.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias*. *The Quarterly Journal of Economics*.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap*. *The Quarterly Journal of Economics*, 125(3):1101–1144.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2015). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, 41(4).

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446.

Davidson, J., Monticini, A., and Peel, D. (2007). Implementing the wild bootstrap using a two-point distribution. *Economics Letters*, 96(3):309 – 315.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2):158–165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, XLII(3):528–554.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, 95(3):932–945.

Downey, D. B. and Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, 77(4):267–282.

Egalite, A. J., Kisida, B., and Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44 – 52.

Ehrenberg, R., Goldhaber, D., and Brewer, D. (1995). Do teachers? race, gender, and ethnicity matter? evidence from the national education longitudinal study of 1988. *Industrial and Labor Relations Review*, 48.

Fairlie, R. W., Hoffmann, F., and Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *The American Economic Review*, 104(8):2567–2591.

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? the effect of student/teacher demographic match on teacher expectations. *Economics of Education Review*, 52:209 – 224.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hazlett, C. (2016). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Hess, F. and L. Leal, D. (1997). Minority teachers, minority students, and college matriculation: a new look at the role-modeling hypothesis. *Policy Studies Journal - POLICY STUD J*, 25:235–248.

Hilmer, C. and Hilmer, M. (2007). Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs. *American Economic Review*, 97(2):422–426.

Hoffmann, F. and Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *The Journal of Human Resources*, 44(2):479–494.

Holt, S. B. and Gershenson, S. (2017). The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, 0(0).

Huber, M., Camponovo, L., Bodory, H., and Lechner, M. (2016). A wild bootstrap algorithm for propensity score matching estimators. FSES Working Papers 470, Faculty of Economics and Social Sciences, University of Freiburg/Fribourg Switzerland.

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. 175:1–21.

Iacus, S., King, G., and Porro, G. (2009). cem: Software for coarsened exact matching. *Journal of Statistical Software, Articles*, 30(9):1–27.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Joensen, J. S. and Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*, 126(593):1129–1163.

Kallus, N. (2016). Generalized optimal matching methods for causal inference.

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching.

King, G., Nielsen, R., Coberley, C., Pope, J. E., and Wells, A. (2011). Comparative effectiveness of matching methods for causal inference.

Krueger, A. (2017). Where have all the workers gone? an inquiry into the decline of the u.s. labor force participation rate. *Brookings Papers on Economic Activity*, 48(2 (Fall)):1–87.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of Public Economics*, 92(10):2083 – 2105.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263 – 279.

Maria Villegas, A., Strom, K., and Lucas, T. (2012). Closing the racial/ethnic gap between students of color and their teachers: An elusive goal. *Equity & Excellence in Education*, 45:283–301.

Mechtenberg, L. (2009). Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages. *Review of Economic Studies*, 76(4):1431–1459.

Neumark, D. and Gardecki, R. (1998). Women helping women? role model and mentoring effects on female ph.d. students in economics. *The Journal of Human Resources*, 33(1):220–246.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732.

Ouazad, A. (2014). Assessed by a teacher like me: Race and teacher assessments. *Education Finance and Policy*, 9(3):334–372.

Ouazad, A. and Page, L. (2012). Students' Perceptions of Teacher Biases: Experimental Economics in Schools. CEE Discussion Papers 0133, Centre for the Economics of Education, LSE.

Penner, E. K. (2016). Teaching for all? teach for america's effects across the distribution of student achievement. *Journal of Research on Educational Effectiveness*, 9(3):259–282.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3):97–118.

Steele, C. M. (1997). A threat in the air. how stereotypes shape intellectual identity and performance. *The American psychologist*, 52 6:613–29.

Steele, J. (2003). Children's gender stereotypes about math: The role of stereotype stratification1. *Journal of Applied Social Psychology*, 33(12):2587–2606.

Terrier, C. (2016). Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement. IZA Discussion Papers 10343, Institute of Labor Economics (IZA).

Williams, W. M. and Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences*, 112(17):5360–5365.

Winters, M. A., Haight, R. C., Swaim, T. T., and Pickering, K. A. (2013). The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data. *Economics of Education Review*, 34:69 – 75.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *The Review of Economics and Statistics*, 86(1):91–107.

APPENDIX A. ADDITIONAL MATERIAL FOR CHAPTER 1

Proof of Bootstrap Failure

Recall that

$$\begin{aligned}\mathrm{Var}\left(\sqrt{N_1}\left(\hat{\tau}^{t*}-\hat{\tau}^{t}\right)\mid\mathbf{Z}\right) &= N_1\mathbb{E}\left[(T_N^{t*})^2+(Q_N^{t*})^2+(R_N^{t*})^2\mid\mathbf{Z}\right]\\ &\quad+N_1\mathbb{E}\left[2\left(T_N^{t*}Q_N^{t*}+T_N^{t*}R_N^{t*}+Q_N^{t*}R_N^{t*}\right)\mid\mathbf{Z}\right] \qquad \text{(A.1)}\end{aligned}$$

Consider each part of the above separately. First,

$$\begin{aligned}\mathbb{E}\left[N_1(T_N^{t*})^2\mid\mathbf{Z}\right] &= \frac{1}{N_1}\sum_{i=1}^{N}D_i\left(\mu(1,X_i)-\mu(0,X_i)-\tau^t\right)\\ &= \frac{1}{N_1}\sum_{i:D_i=1}\left(\mu(1,X_i)-\mu(0,X_i)-\tau^t\right)\\ &\to^p (\sigma_2^t)^2 \text{ by the Law of Large Numbers} \qquad \text{(A.2)}\end{aligned}$$

Second, consider Q_N^{t*}. Let $e_i = Y_i - \mu(D_i, X_i)$. Recalling that $m(i)$ returns the index of the single match to unit i,

$$\begin{aligned}\mathbb{E}\left[N_1(Q_N^{t*})^2\mid\mathbf{Z}\right] &= \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(Y_i(1)-\widehat{Y_i(0)}-\mu(1,X_i)+\mu(0,X_i)\right)^2\mid\mathbf{Z}\right]\\ &= \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(e_i-\mu(0,X_{m(i)})-e_{m(i)}+\mu(0,X_i)\right)^2\mid\mathbf{Z}\right]\\ &= \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(e_i^2+e_{m(i)}^2+\mu(0,X_{m(i)})^2+\mu(0,X_i)^2-2\mu(0,X_i)\mu(0,X_{m(i)})\right)\mid\mathbf{Z}\right] \qquad \text{(A.3)}\end{aligned}$$

where the final step follows because the idiosyncratic errors e_i are by definition independent of each other, and independent of $\mu(d,x)$ functions. Now, decompose into

$$\mathbb{E}\left[N_1(Q_N^{t*})^2 \mid \mathbf{Z}\right] = \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(e_i^2 + e_{m(i)}^2\right) \mid \mathbf{Z}\right] + Q_N^R \tag{A.4}$$

where

$$Q_N^R = \mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(\mu(0,X_{m(i)})^2 + \mu(0,X_i)^2 - 2\mu(0,X_i)\mu(0,X_{m(i)})\right) \mid \mathbf{Z}\right] \tag{A.5}$$

First, dealing with (25), we need to convert the sum to cover the entire sample. To do so, note that for control unit j, e_j will show up once for each time $j \in \mathcal{J}_{M(i)}$. By the definition of K_i, this happens K_i times. Thus,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N_1}\sum_{i:D_i=1}\left(e_i^2 + e_{m(i)}^2\right) \mid \mathbf{Z}\right] &= \frac{1}{N_1}\sum_{i=1}^{N}\left(D_i + (1-D_i)K_i\right)\mathbb{E}\left[e_i^2 \mid \mathbf{Z}\right] \\ &\to^p \frac{1}{N_1}\sum_{i=1}^{N}\left(D_i + (1-D_i)K_i\right)\sigma^2(D_i,X_i)\end{aligned} \tag{A.6}$$

Regarding Q_N^R, note that since $\mu(0,X_i) - \mu(0,X_{m(i)})$ is $o_p(N^{-1/2})$ it must be that $\mu(0,X_{m(i)}) - \mu(0,X_i)$ is also $o_p(N^{-1/2})$, so both terms in Q_N^R are $o_p(N^{-1/2})$, and thus $Q_N^R \to^p 0$.

Finally, for R_N^{t*}, note first that $\mathbb{E}\left[R_N^{t*} \mid \mathbf{Z}\right] = 0$. Thus,

$$\begin{aligned}\mathrm{Var}\left(R_N^{t*} \mid \mathbf{Z}\right) &= \frac{1}{N_1^2}\sum_{i=1}^{N} D_i \mathrm{Var}\left((\tau^t - \hat{\tau}^t)\epsilon_i^*\right) \\ &= \frac{1}{N_1}\mathrm{Var}\left(\tau^t - \hat{\tau}^t\right) \\ &= \frac{1}{N_1}O_p(N^{-1/2}) \\ &= O_p(N^{-1/2})\end{aligned} \tag{A.7}$$

Finally, consider the cross product terms. First, $T_N^{t*}R_N^{t*}$:

$$T_N^{t*}R_N^{t*} = \frac{1}{N_1^2}\sum_{i=1}^{N} D_i\left(\mu(1,X_i)\tau^t - \mu(1,X_i)\hat{\tau}^t - \mu(0,X_i)\tau^t + \mu(0,X_i)\hat{\tau}^t - (\tau^t)^2 + \tau^t\hat{\tau}^t\right)$$

Since $\hat{\tau}^t \to^p \tau^t$, it follows immediately that

$$T_N^{t*} R_N^{t*} \to^p \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \left(\mu(1, X_i)\tau^t - \mu(1, X_i)\tau^t - \mu(0, X_i)\tau^t + \mu(0, X_i)\tau^t - (\tau^t)^2 + (\tau^t)^2 \right)$$

$$\to^p 0 \tag{A.8}$$

It is straightforward to verity that $Q_N^{t*} R_N^{t*} \to^p 0$. The proof is nearly identical. Finally, consider $T^{t*} Q_N^{t*}$,

$$\begin{aligned} T^{t*} Q_N^{t*} = \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \Big(& \mu(1, X_i) Y_i(1) - \mu(1, X_i)\widehat{Y_i(0)} - \mu(1, X_i)^2 + \mu(1, X_i)\mu(0, X_i) \\ & - \mu(0, X_i) Y_i(1) + \mu(0, X_i)\widehat{Y_i(0)} + \mu(0, X_i)\mu(1, X_i) - \mu(0, X_i)^2 \\ & - \tau^t Y_i(1) + \tau^t \widehat{Y_i(0)} + \tau^t \mu(1, X_i) - \tau^t \mu(0, X_i) \Big) \end{aligned} \tag{A.9}$$

Recalling that $Y_i(1) = \mu(1, X_i) + e_i$ and making the substitution, a number of terms cancel, leaving

$$\begin{aligned} T_N^{t*} Q_N^{t*} = \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \Big(& \mu(1, X_i) e_i - \mu(1, X_i)\widehat{Y_i(0)} - \mu(0, X_i) e_i + \mu(0, X_i)\widehat{Y_i(0)} \\ & + \mu(0, X_i)\mu(1, X_i) - \mu(0, X_i)^2 - \tau^t e_i + \tau^t \widehat{Y_i(0)} - \tau^t \mu(0, X_i) \Big) \end{aligned} \tag{A.10}$$

Similarly, recall that $\widehat{Y_i(0)} = Y_{m(i)} = \mu(0, X_{m(i)}) + e_{m(i)}$, and make the substitution.

$$\begin{aligned} T_N^{t*} Q_N^{t*} = \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \Big(& \mu(1, X_i) e_i - \mu(1, X_i)\mu(0, X_{m(i)}) - \mu(1, X_i) e_{m(i)} - \mu(0, X_i) e_i \\ & + \mu(0, X_i)\mu(0, X_{m(i)}) + \mu(0, X_i) e_{m(i)} + \mu(0, X_i)\mu(1, X_i) \\ & - \mu(0, X_i)^2 - \tau^t e_i + \tau^t \mu(0, X_{m(i)}) + \tau^t e_{m(i)} - \tau^t \mu(0, X_i) \Big) \end{aligned} \tag{A.11}$$

Note that when I take probability limits, every term involving e_i or $e_{m(i)}$ goes to zero, so I will remove them now for convenience.

$$\begin{aligned} T_N^{t*} Q_N^{t*} = \frac{1}{N_1^2} \sum_{i=1}^{N} D_i \Big(& \mu(1, X_i) \left[\mu(0, X_i) - \mu(0, X_{m(i)} \right] - \mu(0, X_i) \left[\mu(0, X_i) - \mu(0, X_{m(i)} \right] \\ & - \tau^t \left[\mu(0, X_i) - \mu(0, X_{m(i)} \right] \Big) \end{aligned} \tag{A.12}$$

Since $\mu(0, X_i) - \mu(0, X_{m(i)})$ is $o_p(N^{-1/2})$ when bias correction is unnecessary (Abadie and Imbens, 2006), $T^{t*} N Q_N^{t*} \to^p 0$.

ATC Simulations

The proposed bootstrap trivially works when estimating the ATC using the Abadie and Imbens (2008) DGP, as shown in Figure A.1.

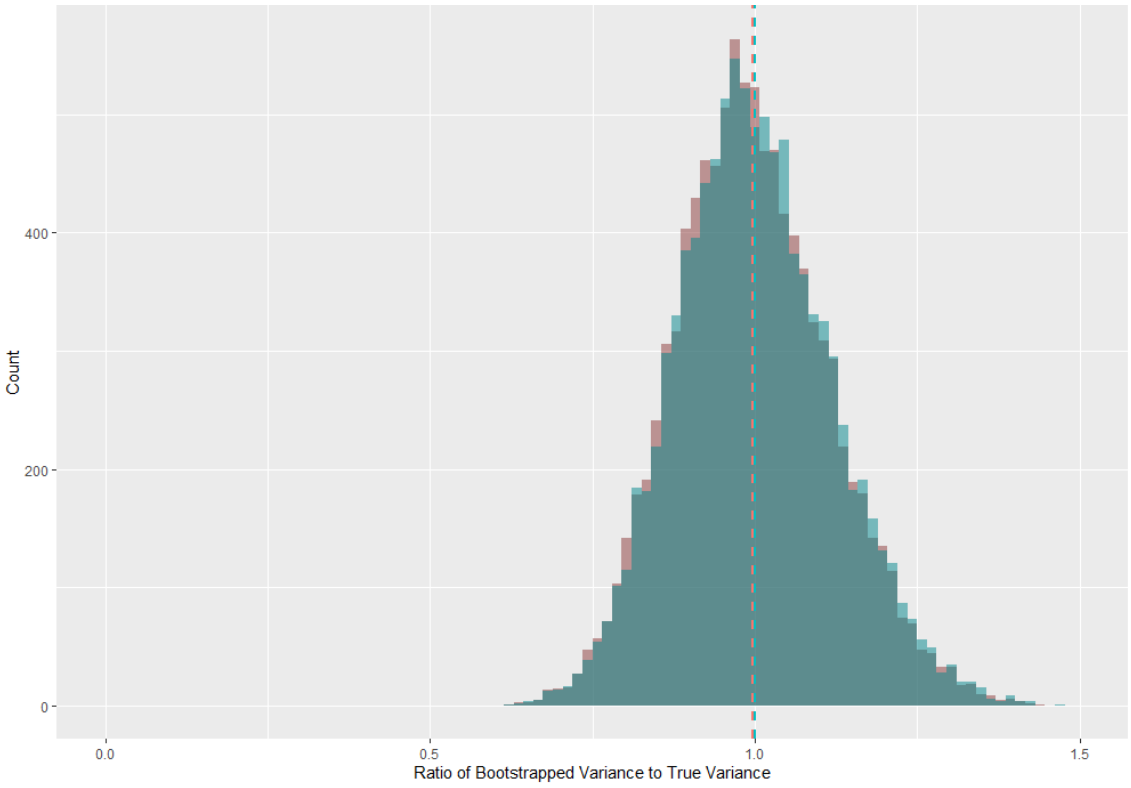

Figure A.1 Proposed and Synthetically Corrected Bootstrap (ATC)

The small differences between the proposed and synthetically corrected bootstrap result from rounding errors when simulating the two procedures. Intuitively, the reason the procedure works in this case is that $e_i = Y_i - \mu(D_i, X_i)$ is uniformly 0 for all treated units, since $Y_i(1)$ has a degenerate distribution. Thus, the value of $\sigma^2(1, X_i)$ is universally 0, so it does not matter that the bootstrap incorrectly weights $\sigma^2(1, X_i)$.

APPENDIX B. ADDITIONAL MATERIAL FOR CHAPTER 2

Proofs

Proof of Theorem 1

Recall that we seek to solve,

$$\min_{k_i} \mathbb{E}\left[\left(\overline{\mu(X_{D_1},0)} - \sum_{D_i=0} k_i Y_i\right)^2\right] \tag{B.1}$$

It is convenient to assume, without loss of generality, that the N_0 control units are indexed by $i = 1, ..., N_0$ and the N_1 treated units by $i = N_0 + 1, ..., N$. Thus, in particular, $\sum_{i=1}^{N_0}$ is equivalent to $\sum_{D_i=0}$. Expanding the square above results in,

$$\min_{k_i} \mathbb{E}\left[\left(\sum_{i=1}^{N_0} k_i Y_i\right)^2 - 2\overline{\mu(X_{D_1},0)}\sum_{i=1}^{N_0} k_i Y_i + \overline{\mu(X_{D_1},0)}^2\right] \tag{B.2}$$

First, consider the leading squared sum:

$$\begin{aligned}\left(\sum_{i=1}^{N_0} k_i Y_i\right)^2 &= \sum_{i=1}^{N_0} k_i^2 Y_i^2 + \sum_{i=1}^{N_0}\sum_{j\neq i} k_i k_j Y_i Y_j \\ &= \sum_{i=1}^{N_0} k_i^2 \left(\mu(X_i,0) + \varepsilon_i\right)^2 + \sum_{i=1}^{N_0}\sum_{j\neq i} k_i k_j \left(\mu(X_i,0) + \varepsilon_i\right)\left(\mu(X_j,0) + \varepsilon_j\right)\end{aligned} \tag{B.3}$$

Since ε_i is mean-zero and independent of ε_j, after taking expectations the minimization problem becomes,

$$\begin{aligned}\min_{k_i}\Bigg[&\sum_{i=1}^{N_0} k_i^2\left(\mu(X_i,0)^2 + \sigma_i^2\right) + \sum_{i=1}^{N_0}\sum_{j\neq i} k_i k_j \mu(X_i,0)\mu(X_j,0) \\ &-2\overline{\mu(X_{D_1},0)}\sum_{i=1}^{N_0} k_i \mu(X_i,0) + \overline{\mu(X_{D_1},0)}^2\Bigg]\end{aligned} \tag{B.4}$$

And the solution can be found by solving a system of N_0 first order conditions:

$$2k_i\left(\mu(X_i,0)^2+\sigma_i^2\right)+2\mu(X_i,0)\sum_{j\neq i}k_j\mu(X_j,0)-2\mu(X_i,0)\overline{\mu(X_{D_1},0)}=0 \tag{B.5}$$

Adopting the convention that a subscript of $-i$ refers to all indices except i, we can re-arrange to arrive at,

$$k_i\left(k_{-i}\right)=\frac{\overline{\mu(X_{D_1},0)}\mu(X_i,0)-\mu(X_i,0)\sum_{j\neq i}k_j\mu(X_j,0)}{\mu(X_i,0)^2+\sigma_i^2} \tag{B.6}$$

Plug this into the first order condition for some $l\neq i$, to find,

$$\begin{aligned}k_l\left(\mu(X_l,0)^2+\sigma_l^2\right)=\overline{\mu(X_{D_1},0)}\mu(X_l,0)-\left(\mu(X_l,0)\sum_{t\neq l,i}k_t\mu(X_t,0)\right)\\-\mu(X_l,0)\mu(X_i,0)\frac{\overline{\mu(X_{D_1},0)}\mu(X_i,0)-\mu(X_i,0)\sum_{j\neq i}k_j\mu(X_j,0)}{\mu(X_i,0)^2+\sigma_i^2}\end{aligned} \tag{B.7}$$

Note that the index $j\neq i$ includes l, and thus:

$$\begin{aligned}k_l\left(\mu(X_l,0)^2+\sigma_l^2\right)=\overline{\mu(X_{D_1},0)}\mu(X_l,0)-\mu(X_l,0)\sum_{t\neq l,i}k_t\mu(X_t,0)\\+\frac{\mu(X_i,0)^2\mu(X_l,0)}{\mu(X_i,0)^2+\sigma_i^2}\sum_{j\neq l,i}k_j\mu(X_j,0)\\-\frac{\overline{\mu(X_{D_1},0)}\mu(X_i,0)^2\mu(X_l,0)}{\mu(X_i,0)^2+\sigma_i^2}+\frac{\mu(X_i,0)^2\mu(X_l,0)^2k_l}{\mu(X_i,0)^2+\sigma_i^2}\end{aligned} \tag{B.8}$$

Since $t\neq l,i$ and $j\neq l,i$ describe the same set, this simplifies to,

$$\begin{aligned}k_l\left(\mu(X_l,0)^2+\sigma_l^2\right)=\overline{\mu(X_{D_1},0)}\mu(X_l,0)-\frac{\mu(X_l,0)\sigma_i^2}{\mu(X_i,0)^2+\sigma_i^2}\sum_{j\neq l,i}k_j\mu(X_j,0)\\-\frac{\overline{\mu(X_{D_1},0)}\mu(X_i,0)^2\mu(X_l,0)}{\mu(X_i,0)^2+\sigma_i^2}+\frac{\mu(X_i,0)^2\mu(X_l,0)^2k_l}{\mu(X_i,0)^2+\sigma_i^2}\end{aligned} \tag{B.9}$$

Bringing the right-hand side of the above over a common denominator and simplifying gives,

$$k_l\left(\mu(X_l,0)^2+\sigma_l^2-\frac{\mu(X_i,0)^2\mu(X_l,0)^2}{\mu(X_i,0)^2+\sigma_i^2}\right)=\frac{\mu(X_l,0)\sigma_i^2}{\mu(X_i,0)^2+\sigma_i^2}\left(\overline{\mu(X_{D_1},0)}-\sum_{j\neq l,i}k_j\mu(X_j,0)\right)$$

Doing the same to the left-hand side gives,

$$k_l = \frac{\mu(X_l,0)\sigma_i^2}{\mu(X_l,0)^2\sigma_i^2 + \mu(X_i,0)^2\sigma_l^2 + \sigma_i^2\sigma_l^2}\left(\overline{\mu(X_{D_1},0)} - \sum_{j\neq l,i} k_j\mu(X_j,0)\right) \tag{B.10}$$

Which allows for a simplified relationship between k_i and k_j to be found:

$$k_j = k_j\frac{X_j}{X_i}\frac{\sigma_i^2}{\sigma_j^2} \tag{B.11}$$

Using this relationship delivers,

$$k_i\left(\mu(X_i,0)^2 + \sigma_i^2\right) = \overline{\mu(X_{D_1},0)}\mu(X_i,0) - \mu(X_i,0)k_i\frac{\sigma_i^2}{\mu(X_i,0)}\sum_{j\neq i}\frac{\mu(X_j,0)^2}{\sigma_j^2}$$

$$k_i\left(\mu(X_i,0)^2 + \sigma_i^2 + \sigma_i^2\sum_{j\neq i}\frac{\mu(X_j,0)^2}{\sigma_j^2}\right) = \overline{\mu(X_{D_1},0)}\mu(X_i,0) \tag{B.12}$$

Noting that the common denominator on the left-hand side is $\prod_{j\neq i}\sigma_j^2$, this becomes,

$$k_i\left(\mu(X_i,0)^2 + \sigma_i^2 + \sigma_i^2\sum_{j\neq i}\left[\frac{\mu(X_j,0)^2}{\sigma_j^2}\frac{\prod_{l\neq j,i}\sigma_l^2}{\prod_{l\neq j,i}\sigma_l^2}\right]\right) = \overline{\mu(X_{D_1},0)}\mu(X_i,0) \tag{B.13}$$

Move the σ_i^2 and σ_j^2 terms into the appropriate product operators to produce,

$$k_i\left(\frac{\left(\mu(X_i,0)^2 + \sigma_i^2\right)\prod_{l\neq i}\sigma_l^2 + \sum_{j\neq i}\mu(X_j,0)^2\prod_{l\neq j}\sigma_l^2}{\prod_{l\neq i}\sigma_l^2}\right) = \overline{\mu(X_{D_1},0)}\mu(X_i,0) \tag{B.14}$$

Finally, note that in the numerator above, the first term can be moved into the summation,

$$k_i = \overline{\mu(X_{D_1},0)}\mu(X_i,0)\frac{\prod_{j\neq i}\sigma_j^2}{\sum_{i=1}^{N_0}\left(\mu(X_i,0)^2\prod_{j\neq i}\sigma_j^2\right) + \prod_{i=1}^{N_0}\sigma_i^2} \tag{B.15}$$

completing the proof.

Proof of Theorem 2

We seek to solve the same minimization problem, but subject to the constraints that $k_i \geq 0$ for all i and $\sum_{i=1}^{N_0} k_i = 1$. The associated Lagrangian is,

$$\mathcal{L} = \sum_{i=1}^{N_0} k_i^2\sigma_i^2 + \left(\sum_{i=1}^{N_0} k_iY_i\right)^2 + \overline{\mu(X_{D_1},0)}^2 - 2\overline{\mu(X_{D_1},0)}\sum_{i=1}^{N_0} k_iY_i - \lambda\left(\sum_{i=1}^{N_0} k_i - 1\right) \tag{B.16}$$

and the general first order condition is,

$$\frac{\partial \mathcal{L}}{\partial k_i} = 2k_i\sigma_i^2 + 2\left(\sum_{i=1}^{N_0} k_i Y_i\right) Y_i - 2\overline{\mu(X_{D_1}, 0)} Y_i = 0 \tag{B.17}$$

Using the first-order condition for k_i and k_1 and solving for the relationship between them gives,

$$k_i = \frac{\sigma_1^2}{\sigma_i^2} k_1 + \left(\sum_{i=1}^{N_0} k_i Y_i - \overline{\mu(X_{D_1}, 0)}\right) \frac{Y_1 - Y_i}{\sigma_i^2} \tag{B.18}$$

Thus:

$$\begin{aligned}\sum_{i=1}^{N_0} k_i &= \sum_{i=1}^{N_0} \frac{\sigma_1^2}{\sigma_i^2} k_1 + \sum_{i=1}^{N_0} \left(\sum_{i=1}^{N_0} k_i Y_i - \overline{\mu(X_{D_1}, 0)}\right) \frac{Y_1 - Y_i}{\sigma_i^2} \\ &= k_1\sigma_1^2 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} + \left(\sum_{i=1}^{N_0} k_i Y_i - \overline{\mu(X_{D_1}, 0)}\right) \left(Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right)\end{aligned} \tag{B.19}$$

From here, it is straightforward to recover,

$$\sum_{i=1}^{N_0} k_i Y_i = k_1\sigma_1^2 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} + \left(\sum_{i=1}^{N_0} k_i Y_i - \overline{\mu(X_{D_1}, 0)}\right) \left(Y_1 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2}\right) \tag{B.20}$$

Subtracting $\overline{\mu(X_{D_1}, 0)}$ from both sides and simplifying delivers,

$$\left(\sum_{i=1}^{N_0} k_i Y_i - \overline{\mu(X_{D_1}, 0)}\right) = \frac{k_1\sigma_1^2 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} - \overline{\mu(X_{D_1}, 0)}}{1 - Y_1 \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} + \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2}} \tag{B.21}$$

Substituting this result into the equation for $\sum k_i$,

$$\sum_{i=1}^{N_0} k_i = k_1\sigma_1^2 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} + \frac{k_1\sigma_1^2 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} - \overline{\mu(X_{D_1}, 0)}}{1 - Y_1 \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} + \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2}} \left(Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right) \tag{B.22}$$

Solving the above for k_1 produces,

$$k_1 = \frac{1 - Y_1 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} + \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} + \overline{\mu(X_{D_1}, 0)} \left(Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right)}{\sigma_1^2 \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} + \left(\sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2}\right) \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2}\right) - \left(\sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right)^2\right)} \tag{B.23}$$

And plugging this into the formula for k_i, after simplifying,

$$\begin{aligned}k_i = &\frac{1 - Y_1 \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2} + \sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2} + \overline{\mu(X_{D_1}, 0)} \left(Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right)}{\sigma_i^2 \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} + \left(\sum_{i=1}^{N_0} \frac{Y_i^2}{\sigma_i^2}\right) \left(\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2}\right) - \left(\sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}\right)^2\right)} \\ &+ \frac{\sum_{i=1}^{N_0} \frac{1}{\sigma_i^2}}{Y_1 \sum_{i=1}^{N_0} \frac{1}{\sigma_i^2} - \sum_{i=1}^{N_0} \frac{Y_i}{\sigma_i^2}} \frac{\sigma_1^2 (Y_1 - Y_i)}{\sigma_i^2}\end{aligned} \tag{B.24}$$

completes the proof of Theorem 2.

Proof of Lemma 1

Suppose there exist two weight vectors $\{k_i\}_{i=1}^{N_0}$ and $\{k_i'\}_{i=1}^{N_0}$. Suppose they differ only in their final two elements - $k'_{N_0-1} = k_{N_0-1} - q$, $k_{N_0} = 0$, and $k'_{N_0} = q$. Consider how the difference in MSE between these two weight vectors changes with q. The difference in MSE between $\{k_i\}_{i=1}^{N_0}$ and $\{k_i'\}_{i=1}^{N_0}$ is given by

$$\begin{aligned}\Delta MSE = {} & 2qk_{N_0-1}\sigma^2_{N_0-1} - q^2\left(\sigma^2_{N_0} + \sigma^2_{N_0-1}\right) - q^2\left(\mu^2_{N_0} + \mu^2_{N_0-1} - 2\mu_{N_0}\mu_{N_0-1}\right) \\ & - 2q\left(\mu_{N_0} - \mu_{N_0-1}\right)\sum_{i=1}^{N_0-2} k_i\mu_i + 2q\overline{\mu(X_{D_1},0)}\left(\mu_{N_0-1} - \mu_{N_0}\right) \\ & - 2qk_{N_0-1}\mu_{N_0-1}\left(\mu_{N_0-1} - \mu_{N_0}\right) \qquad \text{(B.25)}\end{aligned}$$

where I have suppressed notation, setting $\mu_i = \mu(X_i, D_i)$. Minimizing ΔMSE over q produces the following first order condition:

$$\begin{aligned}0 = {} & 2k_{N_0-1}\sigma^2_{N_0-1} - 2q\left(\sigma^2_{N_0} - \sigma^2_{N_0-1}\right) - q\left(\mu^2_{N_0} + \mu^2_{N_0-1} - 2\mu_{N_0}\mu_{N_0-1}\right) \\ & - 2\left(\mu_{N_0} - \mu_{N_0-1}\right)\sum_{i=1}^{N_0-2} k_i\mu_i + 2\overline{\mu(X_{D_1},0)}\left(\mu_{N_0} - \mu_{N_0-1}\right) \\ & - 2k_{N_0-1}\mu_{N_0-1}\left(\mu_{N_0} - \mu_{N_0-1}\right) \qquad \text{(B.26)}\end{aligned}$$

It is straightforward to verify that the second order condition is negative, and in fact ΔMSE is strictly concave. The solution to the minimization problem is

$$q^* = \frac{k_{N_0-1}\sigma^2_{N_0-1} - \left(\mu_{N_0} - \mu_{N_0-1}\right)\left(\sum_{i=1}^{N_0-2} k_i\mu_i + k_{N_0-1}\mu_{N_0-1} - \overline{\mu(X_{D_1},0)}\right)}{\sigma^2_{N_0} + \sigma^2_{N_0-1} + \left(\mu_{N_0} - \mu_{N_0-1}\right)^2} \qquad \text{(B.27)}$$

For convenience, let $\sum_{i=1}^{N_0-2} k_i\mu_i + k_{N_0-1}\mu_{N_0-1} - \overline{\mu(X_{D_1},0)} = Bias_{k_i}$. It is the bias from the estimator using the weight vector $\{k_i\}_{i=1}^{N_0}$. Further, let $d = (\mu_{N_0} - \mu_{N_0-1})$.

Since the above minimization problem was unconstrained, it is in theory possible for q^* to be weakly larger than k_{N_0}, which would imply the MSE-minimizing weight vector would

contain a zero or negative weight. $q^* \geq k_{N_0-1}$ implies

$$k_{N_0-1}\sigma^2_{N_0-1} - d \cdot Bias_{k_i} \geq (\sigma^2_{N_0} + \sigma^2_{N_0-1} + d^2)k_2$$
$$-d\left(\sum_{i=1}^{N_0-1} k_i\mu_i - \overline{\mu(X_{D_1}, 0)}\right) \geq \sigma^2_{N_0}k_{N_0-1} + d^2k_{N_0-1}$$

Proceeding casewise, suppose that $d > 0$ and $Bias_{k_i} > 0$. This implies that some convex combination of μ_i for $i \in \{1, ..., N_0 - 1\}$ is larger in value than the true parameter $\overline{\mu(X_{D_1}, 0)}$, and that $\mu_{N_0} > \mu_{N_0-1}$. Recall that assumption A4 guarantees $\overline{\mu(X_{D_1}, 0)}$ lies between the largest and smallest values of μ_i. Suppose the largest value of μ_i is μ_j. If $j \neq N_0$, a trivial MSE decrease can be achieved by shifting some arbitrarily small amount of weight from k_j to k_{N_0}, which would reduce both bias and variance. If $j = N_0$, it *must* be the case[1] that some $\mu_l > \overline{\mu(X_{D_1}, 0)}$ and that $k_l > 0$. Thus, again, a trivial MSE decrease could be found by reducing the weight k_l and shifting it to some other unit with $\mu_i < \overline{\mu(X_{D_1}, 0)}$. Thus, if $d > 0$ and $Bias_{k_i} > 0$, it cannot be that MSE is minimized.

Suppose that $d > 0$ and $Bias_{k_i} \leq 0$. This implies that $\mu_{N_0} > \mu_{N_0-1}$ and that $\{k_i\}$ is an overestimate of the true parameter. In turn, this trivially implies that an MSE reduction can be achieved by setting q to some arbitrarily small number, as doing so will reduce both bias and variance.

Suppose that $d < 0$ and $Bias_{k_i} > 0$. This implies that $\mu_{N_0} < \mu_{N_0-1}$ and that $\{k_i\}$ produces an underestimate of the true parameter. As above, this implies the existence of a trivial MSE reduction from setting q to some arbitrarily small number.

Finally, suppose that $d < 0$ and $Bias_{k_i} < 0$. Mirroring the first case, this implies that a MSE decrease can be found by shifting weight between units other than N_0 and $N_0 - 1$. Thus, $q^* \geq k_{N_0-1}$ either fails to minimize MSE, or violates assumption A4.

I omit the proof that $q^* < 0$, as the proof is nearly identical to the above. One can proceed casewise to verify that $q^* < 0$ implies either a failure to minimize MSE or a violation of assumption A4.

[1] Otherwise, if $\mu_j > \overline{\mu(X_{D_1}, 0)}$ and the reverse holds for all other μ_i, it could not be that $Bias_{k_i} > 0$, due to assumption A4.

Finally, to verify that $q^* \neq 0$, note that $q^* = 0$ occurs if and only if $k_{N_0-1}\sigma^2_{N_0-1} = d \cdot Bias_{k_i}$. This is entirely possible - but for $\{k_i\}$ to be an MSE-minimizing weight vector, this must hold for all possible transfers between all possible units. In other words, since N_0 has been assigned zero weight, it must hold that

$$k_i\sigma_i^2 = (\mu_{N_0} - \mu_i)\, Bias_{k_i} \;\forall\; i \in \{1, ..., N_0 - 1\} \tag{B.28}$$

This is only possible if N_0 satisfies either $\mu_{N_0} > \mu_i$ or $\mu_{N_0} < \mu_i$ for all $i \in \{1, ..., N_0 - 1\}$ - otherwise $\mu_{N_0} - \mu_i$ would switch signs for at least one such i. Suppose that $\mu_{N_0} > \mu_i$ for all $i \in \{1, ..., N_0 - 1\}$. For the above condition to hold, it must be that $Bias_{k_i}$ is also positive for all units with positive weight, or $\mu_{N_0} > \mu_i$ for all $i \in \{1, ..., N_0 - 1\}$.

First, if $\overline{\mu(X_{D_1}, 0)}$ lies below μ_{N_0} but above all other μ_i, it is not possible for $Bias_{k_i} > 0$ and $k_{N_0} = 0$ to hold simultaneously. Thus, there must be some $\mu_i > \overline{\mu(X_{D_1}, 0)}$ which also satisfies $\mu_i < \mu_{N_0}$. Call this unit j. By setting $k'_j\mu_j + k'_{N_0}\mu_{N_0} = k_j\mu_j$, bias is reduced. Further, since $k'_j + k'_{N_0} < k_j$, 'slack' is introduced to the weight vector that can be allocated to some other k_i. If σ_j^2 and $\sigma^2_{N_0}$ are equivalent or if $\sigma^2_{N_0} < \sigma_j^2$ this change immediately reduces variance while holding bias constant. If $\sigma_j^2 < \sigma^2_{N_0}$, setting $k'_j\mu_j + k'_{N_0}\mu_{N_0} = k_j\mu_j$ may increase the variance of the estimator. In that case, simply reduce k'_{N_0} and deploy the additional slack weight to reduce bias. Since there is some positive k'_{N_0} that reduces the variance of the estimator, it is always possible to find a set of weight shifts that reduces both bias and variance.

In the alternative case, where $\mu_{N_0} < \mu_i$ for all $i \in \{1, ..., N_0 - 1\}$, the proof is again nearly identical, and thus omitted. This completes the proof of Lemma 1.

APPENDIX C. ADDITIONAL MATERIAL FOR CHAPTER 3

Bandwidth Choice

Following Abrevaya et al. (2015), the bandwidth for my estimates was selected as a multiple of the sample standard deviation in the conditioning covariate. I consider four different multipliers - 0.25, 0.5, 1, and 2. While the range of these multipliers is much smaller than that considered by Abrevaya et al. (2015) in their empirical illustration, it will quickly become clear that even the medium bandwidth of 1 causes the CATE estimator to over-smooth to the extent that it becomes no more informative than an ATE estimator.

Recall that my main specification sets the bandwidth multiplier to 0.5. Setting the bandwidth multiplier to 0.25 causes the estimated CATE function to be significantly less smooth (Figures C.1 and C.2). Qualitatively, however, the story is largely unchanged. The worst-performing female students see a negative effect of assignment to a female teacher, while male students see significantly less heterogeneity and no significant negative effects. The effect of reducing the bandwidth multiplier is nearly identical for reading outcomes. The qualitative story of the estimated CATE function is largely unchanged - significant effects are observed in roughly the same places, and the general shape of the function is similar. Again, there appears to be significantly less heterogeneity for male students than for female students.

Moving in the other direction and increasing the bandwidth multiplier pushes the estimated CATE function strongly towards monotonicity, and towards a flat slope (Figure C.3). With a bandwidth multiplier of 1, almost every estimated CATE function is strictly monotonic, and the vast majority of the variation occurs for estimates conditional on the

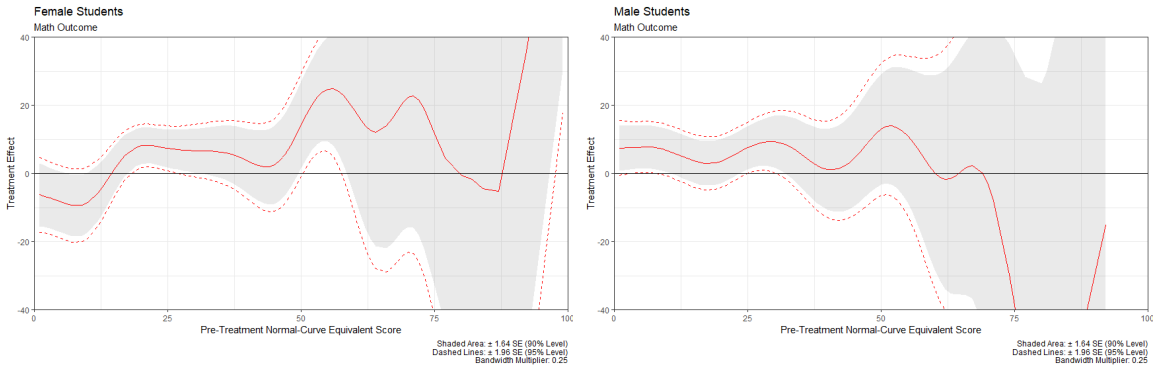

Figure C.1 CATE Estimates (Math) with bandwidth = 0.25

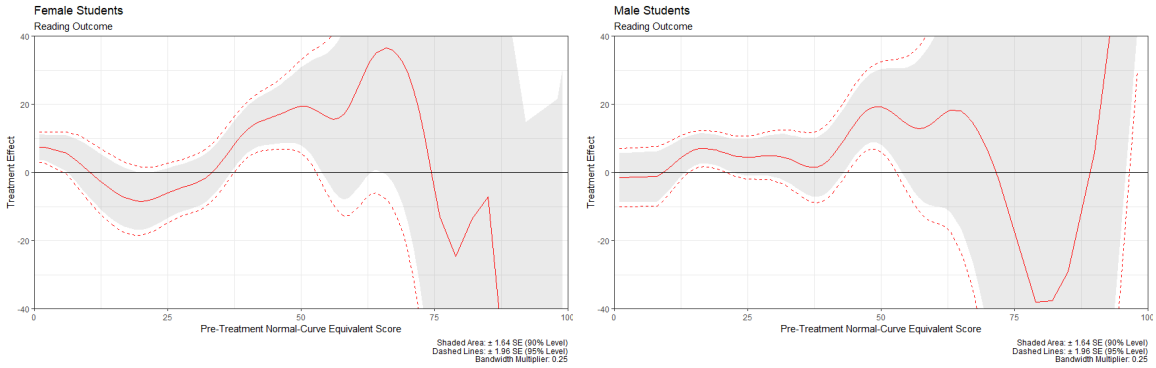

Figure C.2 CATE Estimates (Reading) with bandwidth = 0.25

highest test scores, where very little data is available. Given the heterogeneity present for smaller bandwidths, it seems reasonable to say that at this bandwidth the estimator is clearly over-smoothing. However, note that even with this bandwidth, female students still see notably more heterogeneity than male students in reading, although the difference largely vanishes for math. Increasing the bandwidth multiplier even further to 2 (Figure C.4), forces near-constancy on almost all estimated CATE functions:

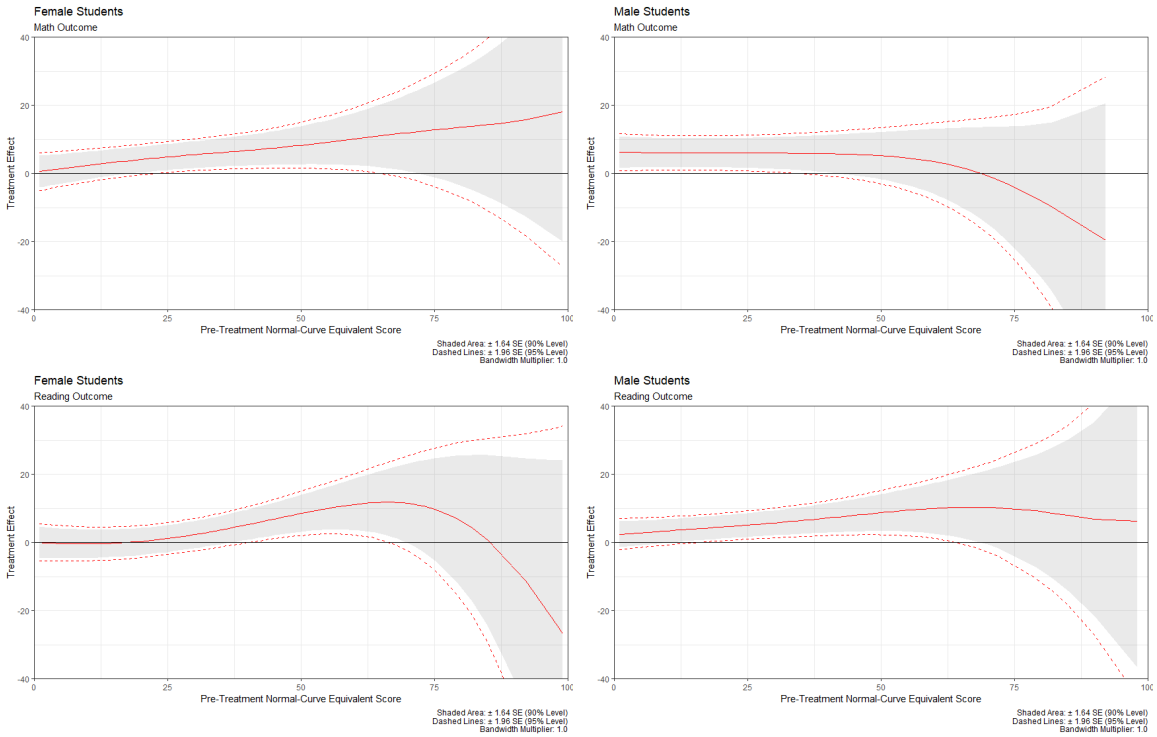

Figure C.3 CATE Estimates with bandwidth = 1

Kernel Choice

As tends to be the case with kernel-based local averaging estimators, the choice of kernel does not have a huge impact on the resulting estimates - bandwidth choice is dramatically more important. I consider two different kernels - the rectangular (uniform) kernel K_r and the Epanechnikov kernel K_e:

$$K_r(u) = \begin{cases} \frac{1}{2} \text{ if } |u| \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

$$K_e(u) = \begin{cases} \frac{3}{4}\left(1 - u^2\right) \text{ if } |u| \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

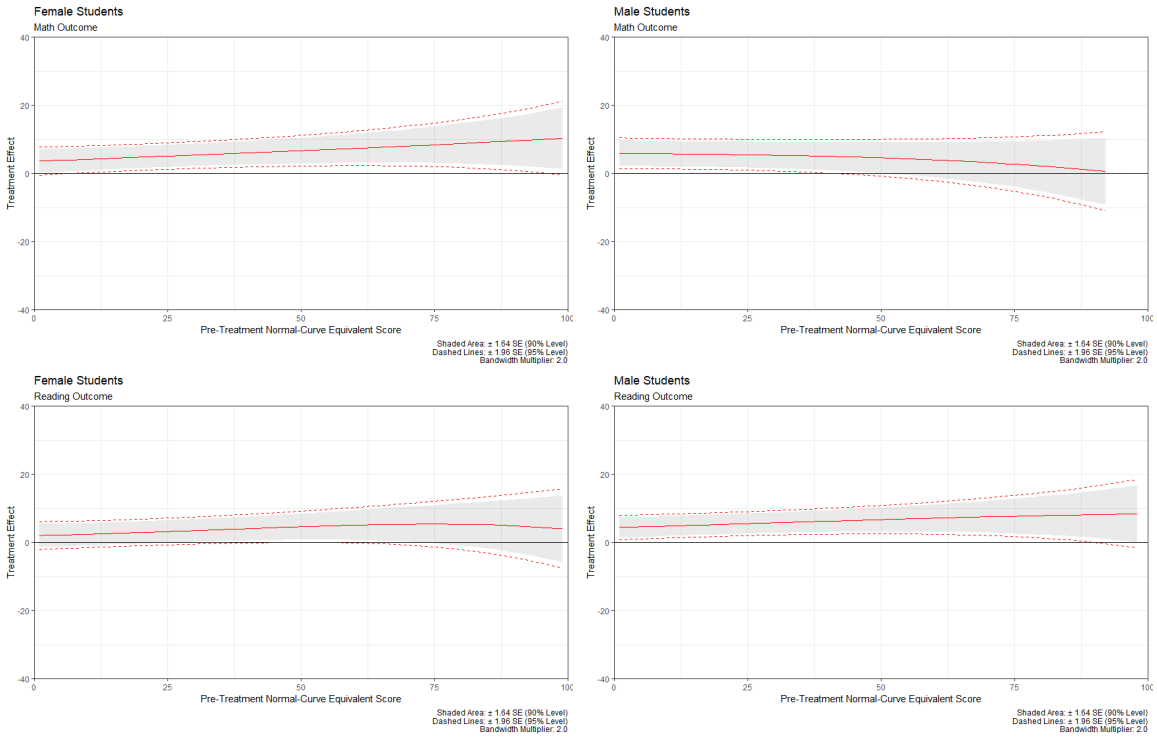

Figure C.4 CATE Estimates with bandwidth = 2

The primary difference between these kernels and the Gaussian kernel is that weights decrease towards zero more rapidly, particularly with the rectangular kernel. This results in less smooth estimates of the CATE function, but the qualitative story is largely unchanged. The effect of bandwidth choice is essentially identical for all kernels, so I report only the results for the intermediate bandwidth multipliers of 0.5 and 1 for these alternative kernels. The effects of other relatively efficient kernels, such as quartic or triweight kernels, are very similar to the effect of the Epanechnikov kernel.

Selection of a rectangular kernel (Figure C.5) generates the least smooth estimates for any given bandwidth. The Epanechnikov kernel (Figure C.6) likewise does not significantly change the qualitative results.

Propensity Score Estimation

Details of the main specification

Recall the main specification:

$$\ln \frac{P(FTEACH_i = 1)}{1 - P(FTEACH_i = 1)} = \beta_0 + \beta_1 SC_i^{'} + \beta_2 TC_i^{'} + \beta_3 R_i^{'} + \beta_4 TFA_i + \beta_5 CS_i + u_i$$

$SC_i^{'}$ is a vector of student characteristics. It includes indicators for a student being black or Hispanic, the relevant pre-treatment test score in math or reading as measured on the normal curve equivalent scale, and an indicator for whether the student's class contained a disruptive student.

$TC_i^{'}$ contains the teacher's experience measured in years as well as indicators for whether the teacher was black or Hispanic. In some of the following alternative specifications, it also includes an indicator for possession of a regular teacher certification.

$R^{'}$ is a vector of region indicators. There were 6 regions in the experiment, containing 7 school districts because the Mississippi Delta contributed two school districts. TFA_i is an indicator for whether the teacher was a TFA teacher or not. CS_i is the class size, measured as the number of students in the class at the end of the year[1].

Alternative specifications

First, I consider the addition of the indicator for a traditional teacher certification (Figure C.7). The main specification excludes this variable because previous research (e.g. Staiger and Rockoff (2010)) suggests that teacher certifications are not good predictors of teacher quality, and thus balancing of samples on teacher certification would be harmful unless such balance could be achieved without cost to balance on another covariate (which is not the case). The results of including teacher certification in the propensity score model largely bear this claim out - the qualitative story is almost identical, and the only real

[1] This is the 'true' class size in that it counts students that are not part of the research sample.

change is an increase in the size of the confidence intervals. This is consistent with the expected effects of including an irrelevant covariate in the propensity score model.

I omit reports for other bandwidth multipliers because the results of that exercise are identical - larger confidence bands, with no significant change to the underlying function.

I also consider a much simpler propensity score specification, dropping the teacher and student demographic variables to leave only pre-test score, class size, teacher experience, and indicators for disrupted class, assignment to a TFA teacher, and region (Figure C.8). While this specification clearly excludes potentially relevant covariates, it also results in a complete elimination of numerically 0 or 1 propensity scores, and far fewer extreme propensity scores. If the effect of student or teacher demographics is limited, this specification may make a profitable bias/variance trade-off. In particular, if sorting of teachers into schools was in fact random, or at least uncorrelated with teacher or school characteristics, this specification would be preferable.

While the results for male students are very marginally consistent with the results from my main specification, particularly in math, it is clear that (as prior research would suggest) the demographic variables excluded in this specification are relevant. If they were irrelevant or had a sufficiently minor impact on outcomes, one would expect to see smaller confidence intervals but a largely similar underlying function from this specification.

Finally, I consider the addition of a school fixed effect to the propensity score model (Figures C.9 and C.10). Since a significant minority of schools contain only female teachers, this causes the trimming behavior to play a larger part in the results - many more students receive propensity scores close to 1 or 0 and are thus subject to the trimming behavior. With my default trimming behavior (setting extreme propensity scores to 0.95 or 0.05), the results are again reasonably similar in terms of qualitative story.

However, for female students in math and male students in reading, these results are no longer robust to changes in the trimming behavior. Dropping students with extreme propensity scores generates the following results

These results suggest that non-TFA teachers are not sorting differentially into schools within a region, which was the only potential source of endogeneity in my main specification. A conservative reading of these robustness checks would suggest that the positive treatment effect I find on male students is potentially uncertain, but conclusions related to the heterogeneity in the effect of teacher gender on students of differing abilities are unaffected.

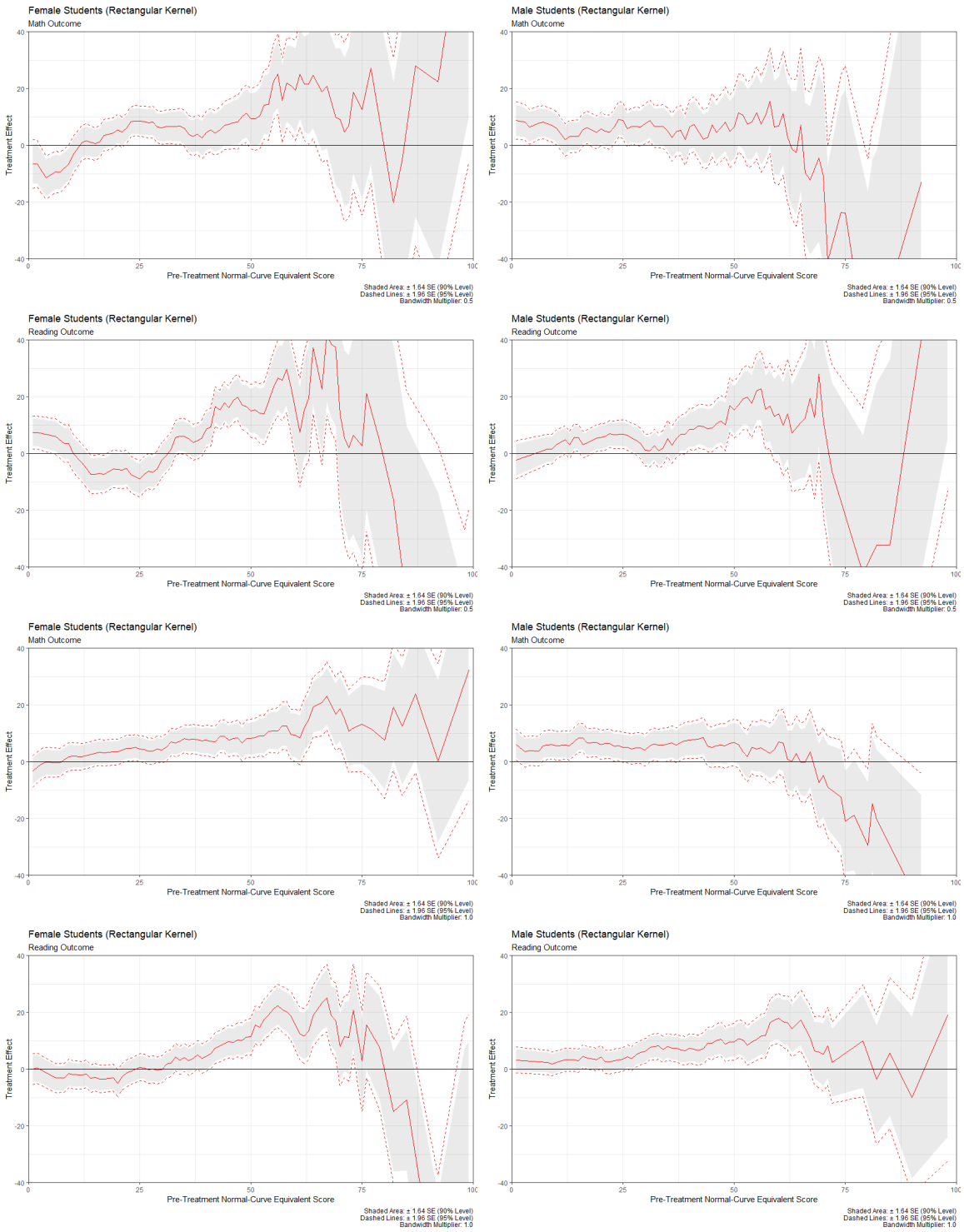

Figure C.5 CATE Estimates with Rectangular Kernel K_r

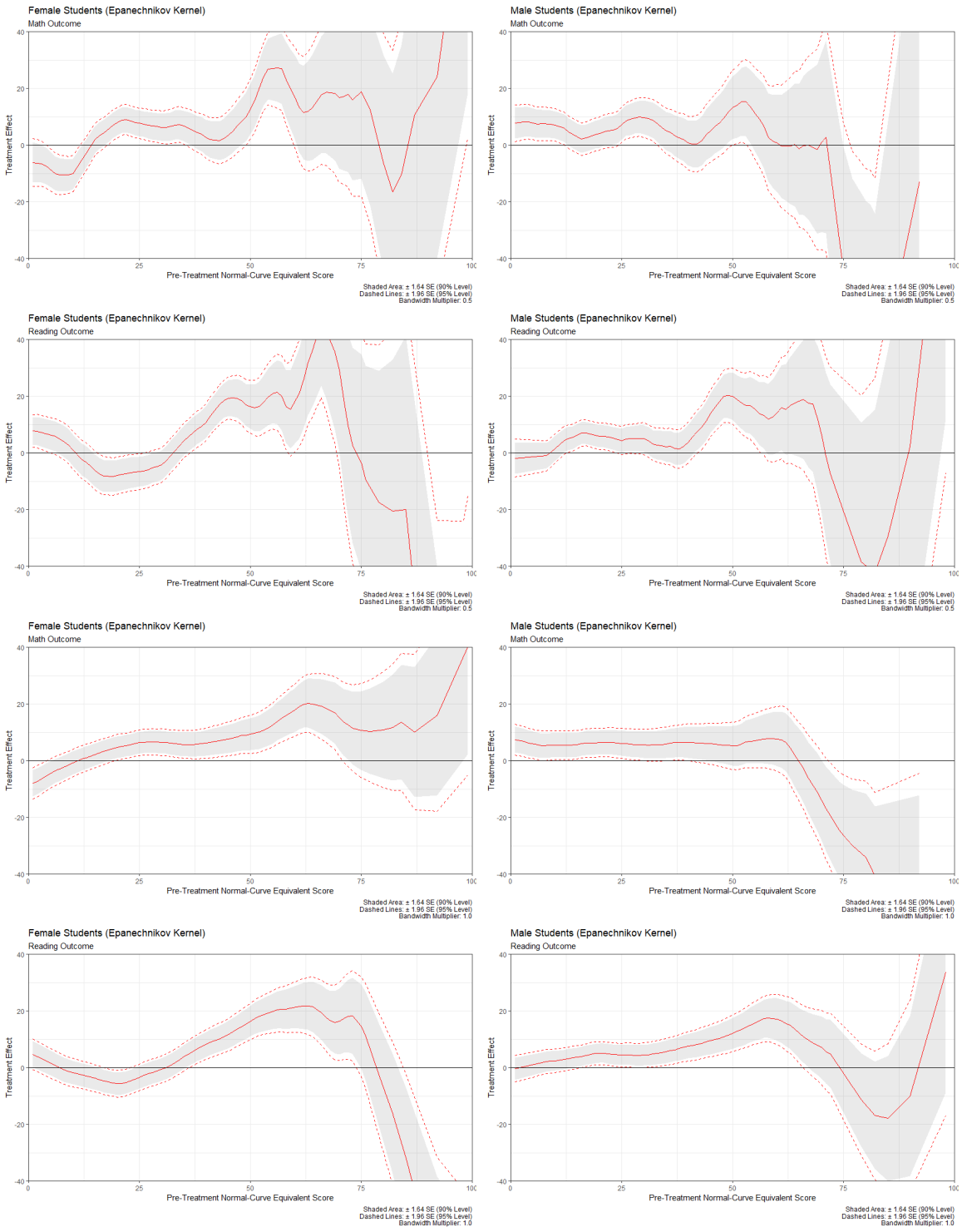

Figure C.6 CATE Estimates with Epanechnikov kernel K_e

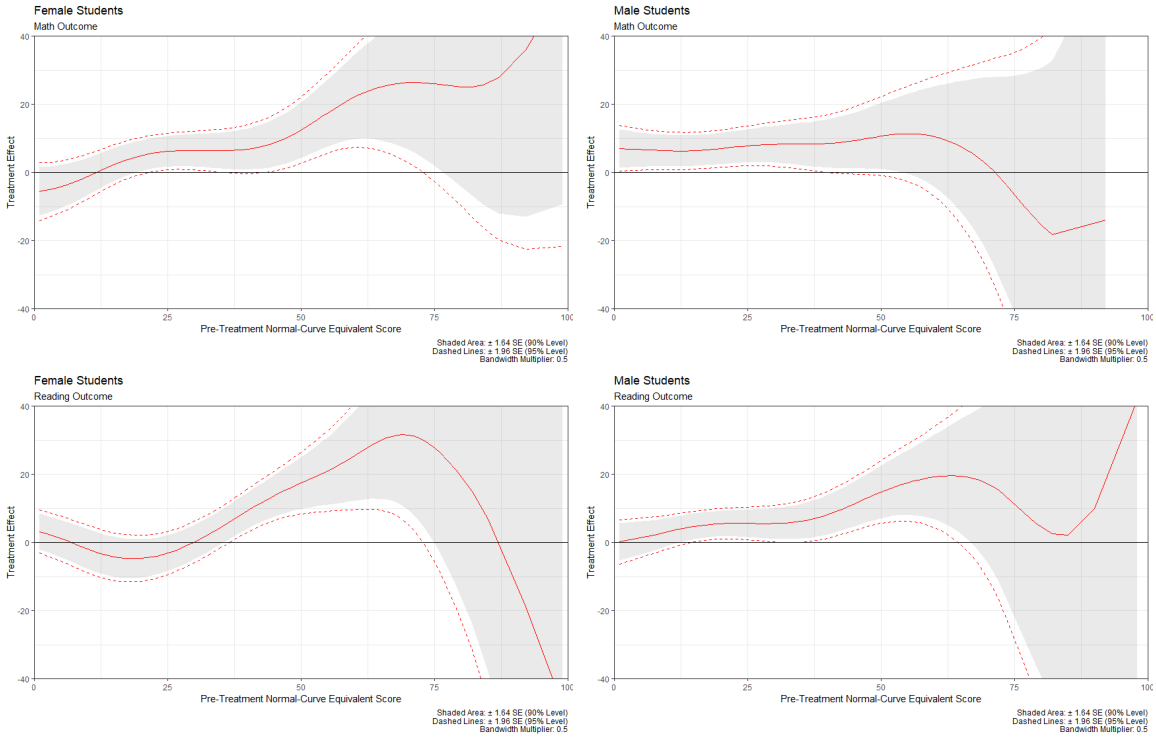

Figure C.7 CATE Estimates with Teacher Certification Indicator

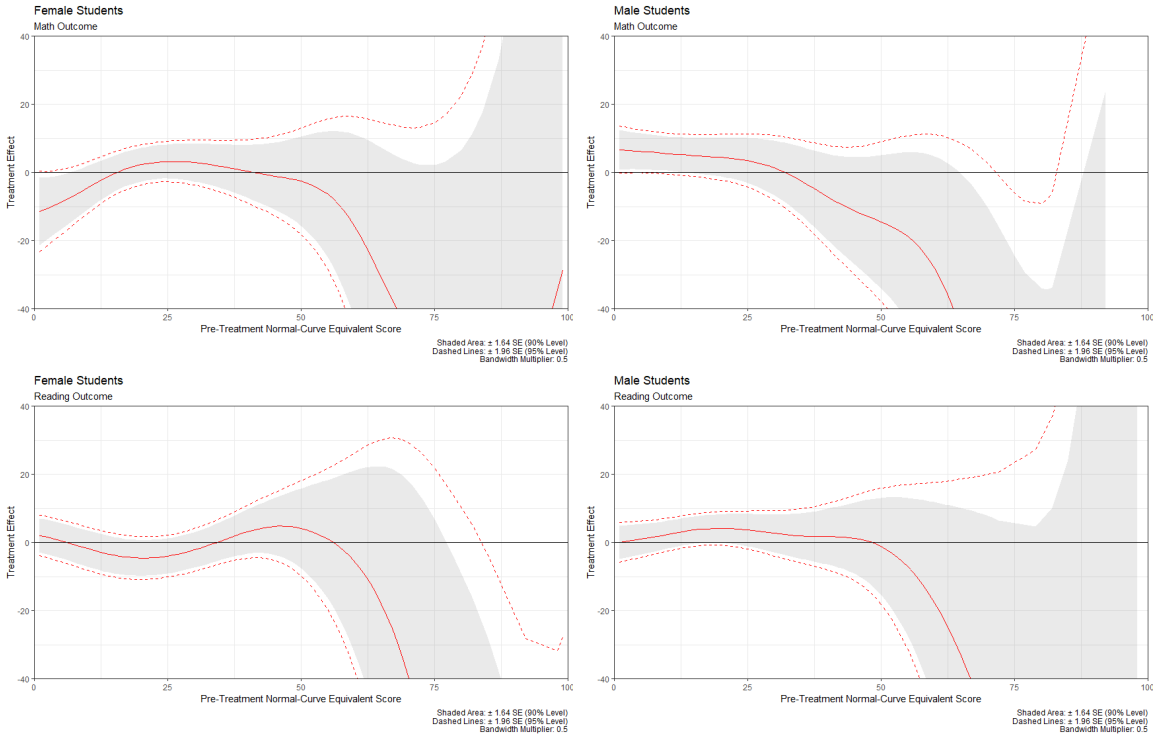

Figure C.8 CATE Estimates without demographics

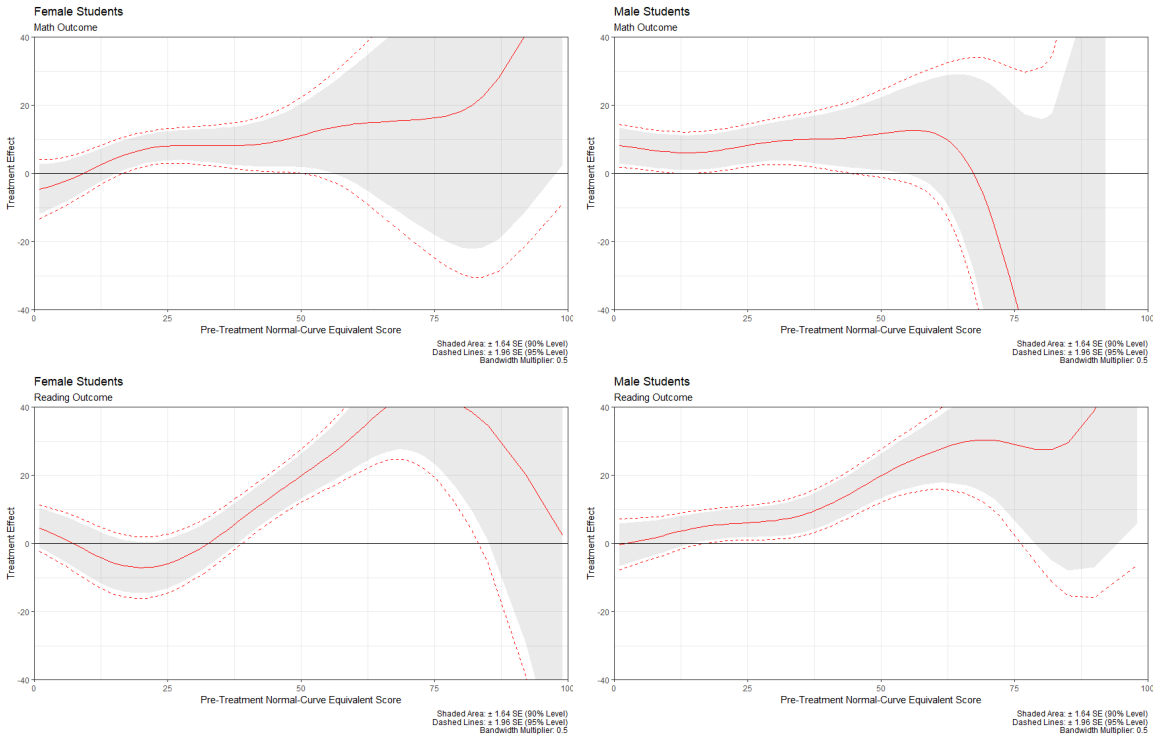

Figure C.9 CATE Estimates (Math) with School Fixed Effects

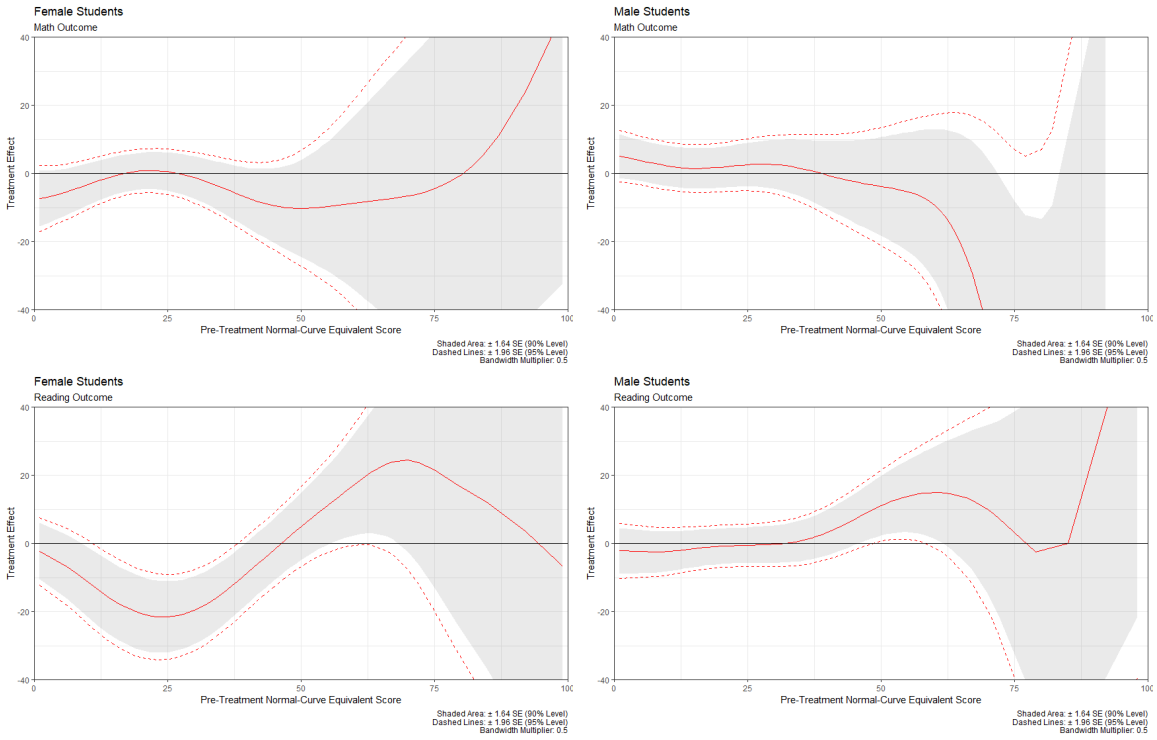

Figure C.10 CATE Estimates (Reading) with School Fixed Effects